Multivariate linear mixed models for statistical genetics

Francesco Paolo Casale

European Bioinformatics Institute
Hughes Hall College
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

October 2016

Declaration of Originality

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as declared in the Preface and specified in the text.

This dissertation does not exceed the specified length limit of 60,000 words as defined by the Biology Degree Committee.

Abstract

In the last decade, genome-wide association studies have helped to advance our understanding of the genetic architecture of many important traits, including diseases. However, the statistical analysis of genotype-phenotype associations remains challenging due to multiple factors. First, many traits have polygenic architectures, which means that they are controlled by a large number of variants with small individual effects. Second, as increasingly deep phenotype data are being generated there is a need for multivariate analysis approaches to leverage multiple related phenotypes while retaining computational efficiency. Additionally, genetic analyses are confronted by strong confounding factors that can create spurious associations when not properly accounted for in the statistical model. We here derive more flexible methods that allow integrating genetic effects across variants and multiple quantitative traits. To do so, we build on the classical linear mixed model (LMM), a widely adopted framework for genetic studies.

The first contribution of this thesis is mtSet, an efficient mixed-model approach that enables genome-wide association testing between sets of genetic variants and multiple traits while accounting for confounding factors. In both simulations and real-data applications we demonstrate that mtSet effectively combines the advantages of variant-set and multi-trait analyses.

Next, we present a new model for gene-context interactions that builds on mtSet. The proposed interaction set test (iSet) yields increased statistical power for detecting polygenic interactions. Additionally, iSet enables the identification of genetic loci that are associated with different configurations of causal variants across contexts. After benchmarking the proposed method using simulated data, we consider two applications to real datasets, where we investigate genetic effects on gene expression across different cellular contexts and sex-specific genetic effects on lipid levels.

Finally, we describe LIMIX, a software framework for the flexible implementation of different LMMs. Most of the models considered in this thesis, including mtSet and iSet, are implemented and available in LIMIX. A unique aspect of the software is an inference framework that allows a large class of genetic models to be defined and, in many cases, to be efficiently fitted by exploiting specific algebraic properties. We demonstrate the utility of this software suite in two applied collaboration projects.

Taken together, this thesis demonstrates the value of flexible and integrative modelling in genetics and contributes new statistical methods for genetic analysis. These approaches generalise previous models, yet retain the computational efficiency that is needed to tackle large genetic datasets.

Acknowledgments

First and foremost, I would like to thank my supervisor Oliver Stegle for being a superb advisor and mentor. His enthusiastic guidance, feedback and support have been invaluable to me, for the work in this thesis and beyond.

I am thankful to EMBL-EBI and the University of Cambridge for providing an excellent and stimulating environment to perform my research. I would also like to thank my advisory committee. A special thank goes to my internal advisor John Marioni for all the stimulating discussions and his continuous presence and support, thanks also to Jan Korbel and Carl Rasmussen for the precious feedback on my work. A big thank also to Ewan Birney for his scientific advice and support.

I am grateful to the whole Stegle group for all the insightful discussions, their feedback and the time together. A special thank goes to the people I had the opportunity to work more closely with: Barbara Rakitsch, Amelie Baud, Rachel Moore and Danilo Horta.

Thanks to the numerous proofreaders of this thesis: Rachel Moore, Davis McCarthy, Hannah Meyer, Nils Koelling, Amelie Baud, Myrto Kostadima, Verena Zuber, Danilo Horta, Tommaso Leonardi, Barbara Rakitsch, Leland Taylor and Na Cai.

Being in Cambridge and at the EBI has been marvellous, I feel lucky and proud to have been part of these communities, of course for the science, but not only that. Many people have contributed to make this experience astonishingly pleasant and rich in so many ways. In particular, thanks to Michele, Myrto, Nils, Mitra, Angela, Ernest, Ana, Maria, Steve, Konrad and Joana for the fantastic time together. A special thank also to Salvatore, Martina and Danilo.

I am grateful to my parents Raffaele and Emma and my brother Ivo for their unconditional love and support. Thanks also to Francesco and Andrea for their love and belief. Last but not least, I would like to thank Federica for being so supportive and patient along these years. There are no words to describe how lucky I feel to have such an exceptional partner in life.

Contents

1 Introduction **1**

1.1 Genetic variation 1

1.1.1 DNA and genetic variants 1

1.1.2 The biological process underlying inheritance 2

1.1.3 Human genetic variation 4

1.2 From genetic variation to phenotype 5

1.2.1 The first genetic models 5

1.2.2 Linkage analysis 6

1.2.3 Genome-wide association studies 7

1.2.4 Molecular mechanisms from genotype to phenotype 8

1.3 Statistical challenges in genetics 10

1.4 Thesis overview and individual contributions 13

2 Linear mixed models for genetic analyses **17**

2.1 Linear regression 17

2.1.1 Maximum likelihood solution 18

2.2 Linear models for genome-wide association studies 20

2.2.1 Statistical hypothesis testing 20

2.2.2 Multiple hypothesis testing correction 21

2.2.3 Distribution of P values and QQ plot 23

2.2.4 Accounting for confounding in the linear model 24

2.3 Genetic analysis with the linear mixed model 25

2.3.1 Accounting for confounding using the linear mixed model 27

2.3.2 Genetic relatedness matrix 28

2.3.3 Efficient linear mixed models for GWAS 31

2.3.4 Variance component models 34

2.3.5 Set tests 36

2.3.6 Genomic predictions . . . 37
2.4 Extension to the analysis of multiple traits . . . 39
2.4.1 Mathematical background . . . 39
2.4.2 The matrix-variate linear mixed model . . . 40
2.4.3 Association testing . . . 42
2.4.4 Efficient implementation . . . 43

3 Efficient set tests for joint analysis of correlated traits 47
3.1 A multi-trait set test . . . 48
3.1.1 The model . . . 48
3.1.2 Statistical testing . . . 51
3.1.3 Efficient parameter inference . . . 53
3.1.4 Analyses of cohorts with unrelated individuals . . . 57
3.1.5 Relationship to existing methods . . . 58
3.2 Simulation study . . . 59
3.2.1 Genotype simulation strategy . . . 59
3.2.2 Phenotype simulation strategy . . . 60
3.2.3 Empirical complexity and scalability . . . 63
3.2.4 Calibration of P values . . . 64
3.2.5 Power comparison . . . 66
3.3 Applications to real data . . . 68
3.3.1 Genetic analysis of lipid traits in human . . . 70
3.3.2 Genetic analysis of haematology traits in rat . . . 71
3.4 Summary and discussion . . . 72

4 Testing for polygenic interactions using set tests 75
4.1 The interaction set test . . . 77
4.1.1 Model derivation . . . 77
4.1.2 Statistical testing . . . 79
4.1.3 Interpretation of the variance parameters . . . 82
4.1.4 Relationship to existing interaction tests . . . 83
4.2 Simulation study . . . 86
4.2.1 Phenotype simulation strategy . . . 86
4.2.2 Illustration case . . . 88
4.2.3 Calibration of P values . . . 90
4.2.4 Power comparison . . . 90
4.2.5 Variance decomposition . . . 93

4.3 Analysis of stimulus-specific eQTLs in monocytes . . . 93
4.3.1 Data preprocessing . . . 94
4.3.2 Mapping of associations and interactions . . . 94
4.3.3 Mechanistic underpinning of heterogeneity eQTLs . . . 96
4.3.4 Note on opposite effects . . . 99
4.4 Extension to analysis of stratified designs . . . 101
4.4.1 Model derivation . . . 102
4.4.2 Simulations . . . 104
4.4.3 Application to gene-by-sex interaction analysis in lipid traits . . . 104
4.5 Summary and discussion . . . 107

5 Flexible LInear MIXed models 109
5.1 A flexible inference framework for linear mixed models . . . 110
5.1.1 Basic classes for inference . . . 110
5.1.2 Covariance models . . . 112
5.1.3 Kronecker-structured covariance models . . . 113
5.1.4 An example of complex covariance model to study social effects . . . 114
5.1.5 Exploiting covariance structures . . . 116
5.1.6 Other flexible inference frameworks . . . 116
5.2 Modules for genetic analyses . . . 117
5.2.1 The variance decomposition module . . . 117
5.2.2 Flexible fixed effect tests in multi-trait mixed models . . . 119
5.3 Vignettes . . . 120
5.3.1 A genetic study of transcription initiation in Drosophila . . . 120
5.3.2 Dissecting the genetic and the epigenetic component of gene expression . . . 123
5.4 Summary and discussion . . . 128

6 Concluding remarks 131

A Derivations 137
A.1 Restricted maximum likelihood . . . 137
A.2 Implementation of LMMs with two-Konecker covariance matrices . . . 138
A.3 Implementation of mtSet gradients . . . 142
A.4 Implementation of mtSet-PC . . . 145
A.5 Implementation of mtSet-LowRankBg . . . 150
A.6 Implementation of iSet for stratification analysis . . . 153

B Supplementary Results for: Efficient set tests for joint analysis of correlated traits **157**
B.1 Supplementary tables . . . 158
B.2 Supplementary figures . . . 169

C Supplementary results for: Testing for polygenic interactions using set tests **181**
C.1 Supplementary tables . . . 181
C.2 Supplementary figures . . . 186

D Supplementary material for: Flexible LINear MIXed models **197**
D.1 Covariance functions and Gaussian processes . . . 197
D.2 Basic covariance models . . . 198
D.3 Standard errors . . . 200
D.4 Supplementary information for the analysis in BluePrint WP10 . . . 200
D.4.1 Molecular assays and data preprocessing . . . 200
D.4.2 Accounting for sample heterogeneity . . . 202
D.4.3 Supplementary Figures . . . 202

E Publications **205**

1 Introduction

The main contributions of this thesis are in the field of statistical genetics. Statistical genetics is concerned with the development and the application of statistical methods to study genetic variation and inheritance in living organisms. In this chapter, I give an overview of this field and the necessary background for the work in this thesis. After discussing genetic variation and its patterns in human in Section 1.1, I give an overview of models and methods to study genetic effects on phenotypes in Section 1.2. In Section 1.3, I discuss some of the statistical challenges in modern genetics. Finally, Section 1.4 concludes this chapter with an outline of the thesis structure and the main contributions.

1.1 Genetic variation

In this section I discuss the mechanisms of genetic variation and inheritance.

1.1.1 DNA and genetic variants

All the information needed to build and sustain an organism is encoded in its genome. The genome is organised into chromosomes, which consist of long molecules of deoxyribonucleic acid (DNA) (Brosius, 2009) and other elements. In chromosomes, DNA is composed of two strands, which are sequences of smaller molecules called nucleotides (Alberts et al., 2014). Each nucleotide in the sequence contains one of four distinct bases (adenine - A, thymine - T, guanine - G or cytosine - C). The two strands are kept together by hydrogen bonds between opposite bases. As these bonds can form only between specific pairs of bases (G with C and A with T) the two sequences are complementary. The number, length and shape of chromosomes differ between species. The human genome consists of 23 pairs of chromosomes: 22 pairs of non-identical copies of autosomal chromosomes (one inherited from the father and the other from the mother) and one pair of sex chromosomes. Females have two non-identical copies

of the X chromosome (one from each of the two parents), while males have only one copy of the X chromosome (from the mother) and one copy of the Y chromosome (from the father). The two non-identical copies of a chromosome are called homologous.

The human genome consists of approximately 3 billion base pairs (bp) (Venter et al., 2001; Lander et al., 2001). Pairs of individuals share on average 99.5% of the DNA sequence (Levy et al., 2007). A change in the sequence across individuals in a population is commonly referred to as a genetic variant while the different forms of the sequence are called alleles. An individual's collection of alleles at a genetic locus (i.e., at a location in the genome) is referred to as the genotype of that individual. An individual is homozygous at a genetic locus on an autosomal chromosome if it has two copies of the same allele; otherwise it is heterozygous at the locus. Depending on the length and type of the sequence variation we can distinguish two types of genetic variants: single nucleotide polymorphisms and structural variants (Frazer et al., 2009). Single nucleotide polymorphisms (SNPs) are substitutions of a single base pair and are the most common type of genetic variation. Most SNPs are bi-allelic, meaning that only two possible alleles are observed in the population (International HapMap Consortium, 2005). Structural variants are changes that involve more than a single base pair and include short insertions and deletions (collectively referred to as indels) as well as larger variants at the chromosome scale, including duplications, deletions, inversions and insertions.

1.1.2 The biological process underlying inheritance

Meiosis is the process of cell division that leads to the formation of sperm cells in males and egg cells in females, collectively called gametes. In contrast to somatic cells, which have two copies of each chromosome (diploid), gametes only have one copy of each chromosome (haploid). The union of a sperm and an egg cell generates a diploid cell, the zygote, which is the first cell of a new organism.

Meiosis starts with the pairing of homologous chromosomes in a diploid cell. Each chromosome is then duplicated giving rise to a pair of chromatids. At the end of this process each chromosome has four homologous chromatids. Identical and non-identical copies of homologous chromatids are called respectively sister and non-sister chromatids. At this stage non-sister chromatids may exchange segments of genetic material, an event that is known as a crossover. Finally, through two rounds of cell division the cell gives rise to four gametes. Each gamete randomly receives one of the four homologous chromatids from each chromosome. This selection occurs independently for each chromosome, leading to a mixture of maternal and paternal chromosomes in

each gamete. An important consequence of crossovers is that gametes may contain chromosomes that are a mixture of the two grandparental chromosomes from the same parent. During the process of fertilisation a sperm and an egg fuse to generate the zygote. Each zygote inherits approximately 1/4 of the genetic material from each of the grandparents. The processes of meiosis and fertilisation are shown in **Fig. 1.1**.

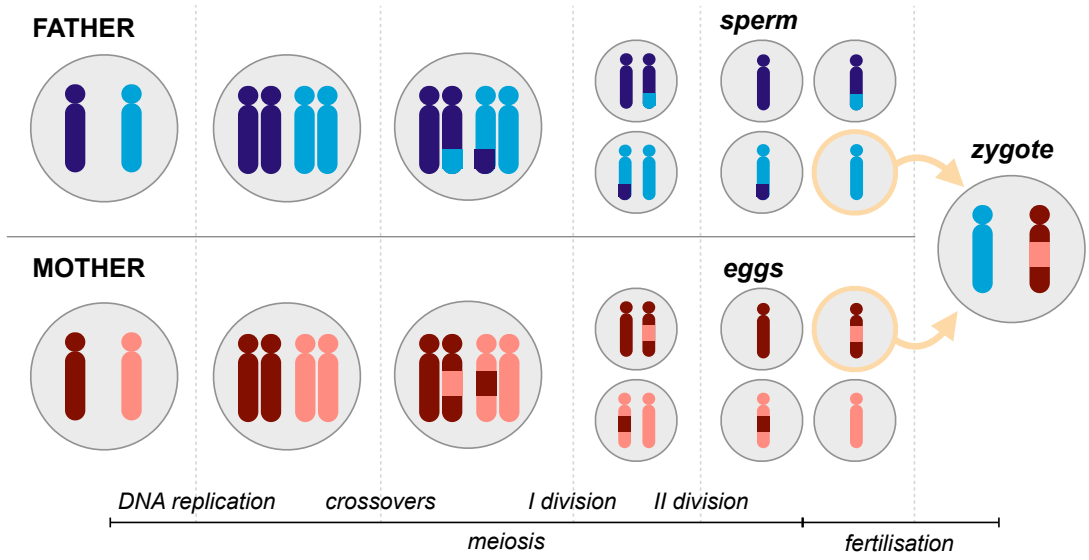


Figure 1.1: **Meiosis and fertilisation.** The figure illustrates key stages of the processes of meiosis and fertilisation, for simplicity considering a single chromosome.

Crossovers are random events and the rate of their occurrence depends on many factors, including sex, chromosome location, temperature and age. The occurrence of a crossover in a region reduces the likelihood of a new crossover in proximity. In every meiotic event, on average approximately 55 crossovers occur in males and 82 in females (Laird and Lange, 2010). The number of crossovers taking place between two loci is not an observable quantity. However, if genotype data are available one can sometimes infer whether a recombination event between two loci has occurred (Morgan, 1915; Laird and Lange, 2010). Note that the occurrence of a recombination only gives information about whether an odd or an even number of crossovers has taken place between the two loci. If at least one crossover between two loci has occurred then the fraction of gametes with a recombination is 50%. The rate of recombination between two loci can thus be estimated as $\theta = (1 - P_0)/2$ (Mather, 1938), where P_0 is the probability of no crossover. For loci that are extremely close on the same chromosome we have $P_0 \approx 1$ and $\theta \approx 0$. These loci tend to be inherited together and are therefore

in linkage. Conversely, for loci that are far apart on the chromosome we have $P_0 \approx 0$ and $\theta = 1$. In linkage studies, recombination rates have been used to infer the distance between loci on chromosomes and thus generate genetic maps (see Section 1.2.2).

1.1.3 Human genetic variation

Allele frequency is an important property of genetic variants. The frequency of an allele at a genetic variant is the proportion of chromosomes in the population that carry that allele (Laird and Lange, 2010). For a bi-allelic variant, the frequency of the less common allele (minor allele) is called the minor allele frequency (MAF). For multi-allelic variants, the MAF typically indicates the frequency of the second most common allele.

Based on their MAF, genetic variants are classified as either common or rare, where commonly a cut-off of 1% is used to define rare variants (MAF$<$1%) and common variants (MAF$>$1%) (Frazer et al., 2009). Comparative analyses of genomic sequences have provided evidence for over ten millions of common variants (Frazer et al., 2009). Importantly, allele frequencies of common variants vary widely across groups of individuals from different geographical regions, with many common variants segregating across populations. This phenomenon is known as population structure and is the consequence of non-random mating primarily due to geographical distance. In contrast to common variants, rare variants are unnumbered and include variants that are specific to close relatives or single individuals (Frazer et al., 2009).

Another central concept in genetic variation is linkage disequilibrium (LD). Two loci are in LD if some combinations of their alleles tend to occur together more of less often than expected by chance. There are different factors that determine LD between variants, including genetic linkage (see above), selection and population structure (Slatkin, 2008). Different measures of LD have been proposed over the years (for a review see Devlin and Risch, 1995). A widely used measure of LD is the squared Pearson correlation between the minor allele counts in the population (Benner et al., 2016; Pickrell et al., 2016; Yang et al., 2012).

The sequencing of the first human genome (Venter et al., 2001; Lander et al., 2001) was followed by the start of the HapMap Project (Gibbs et al., 2003) that aimed at characterising patterns of human genetic variation in different populations with a focus on common variants. This international effort genotyped more than 10 million SNPs in 270 individuals from four human populations (Nigerians, Europeans, Chinese and Japanese). One fundamental result of this study was the haploblock structure of the human genome: the human genome is characterised by regions in which recombina-

tion events are limited, resulting in haplotype blocks[1] characterised by high LD within them. Notably, 550,000 haplotype blocks in Europeans and Asians and 1,100,000 haplotype blocks in Africans were sufficient to accurately summarise the genetic variation described by common SNPs with MAF $> 5\%$ (International HapMap Consortium, 2005). An important implication of this finding is that sparse genotype data at tagging SNPs are sufficient to reconstruct most of the genome-wide genetic variation of common variants, an insight that is key to the success of genome-wide association studies (see Section 1.2.3) and genotype imputation (Howie et al., 2009; Marchini and Howie, 2010).

Exploiting the higher coverage at reduced costs offered by next-generation sequencing (NGS) technologies (Metzker, 2010), the 1000 Genomes Project aimed at studying high-coverage sequences of at least 1,000 individuals from different world-wide populations (1000 Genomes Project Consortium, 2010). In three phases, the consortium generated reference maps of human genetic variation, first considering 1,092 individuals form 14 populations (1000 Genomes Project Consortium, 2012) to finally characterise the genetic variation between 2,504 individuals from 26 human populations. In addition to characterising SNPs and short indels (up to 500 bp, 1000 Genomes Project Consortium, 2015), the 1000 Genomes project has also progressed the boundaries of detecting and genotyping structural variants (1000 Genomes Project Consortium, 2015). Collectively, the results from the project provide "the most comprehensive view of global human variation so far" (Birney and Soranzo, 2015).

1.2 From genetic variation to phenotype

Understanding how genetic variation affects phenotypes is a long standing goal in biology. In this thesis, I will interchangeably use the terms "phenotype" and "trait" to indicate any measurable feature of an individual, which can be a disease state as well as a molecular or cellular feature. Additionally, I will refer to loci that are associated with a trait as a quantitative trait locus (QTL) for that trait. In this section, I give a brief historical overview of the approaches that have been used to study genetic effects on traits.

1.2.1 The first genetic models

Many consider Mendel's work "Experiments in plant hybridisation" the beginning of statistical genetics (Mendel, 1866). Studying inheritance of dichotomous traits in plants,

[1] A haplotype is a set of alleles on the same chromosome.

Mendel discovered the existence of a dominant and a recessive form of a trait. Analysis of phenotype data in successive generations led Mendel to formulate the law of segregation: "One allele of each parent is randomly and independently selected, with probability 1/2, for transmission to the offspring; the alleles unite randomly to form the offspring's genotype" (Laird and Lange, 2010).

In parallel to Mendel's work, Francis Galton, fascinated by Darwin's book "On the Origin of Species" (1859), studied inheritance of height and intelligence in human (Galton, 1869). The traits considered by Galton seemed to follow completely different laws of inheritance from those described in Mendel's work: the parental forms of these traits appeared somehow to be mixed in the offspring. Such traits that do not follow Mendelian patterns of inheritance are commonly referred to as complex traits.

This paradox was solved in 1918 by the theoretical work of the statistician Ronald Fisher (Fisher, 1918). Fisher showed that an additive genetic model with a large number of small-effect loci results in a continuous normally distributed trait. Moreover, he proved that the phenotypic correlation between individuals is proportional to the quantity of genetic material they share. This result demonstrated that heritable complex traits are affected by many genes, each following Mendel's principle of inheritance. A few decades later, building on Fisher's work, Charles Henderson derived the solution of the mixed model equation (Henderson, 1950). Nowadays linear mixed models constitute the standard tool for many genetic analyses and are the basic building block of this thesis.

1.2.2 Linkage analysis

Genetic studies were initially performed solely using phenotype data. In the beginning of the 20th century, the development of the first genotyping technologies introduced the possibility to identify the genomic position of genetic markers and disease genes, giving rise to a branch of statistical genetics known as gene mapping. The first statistical method used for gene mapping was linkage analysis. The core idea underlying this approach is to use the realised recombination rate between pairs of loci in a short pedigree to infer relative genomic distances and build linkage maps (see Griffiths, 2005, chap 5). The same methodology has been successfully applied to locate disease variants for many Mendelian traits by using genotype and phenotype data on pedigrees over one/two generations. The strategy employed to map disease loci is to test for linkage between any of the genotyped markers and the disease variant (whose genotype is inferred from the phenotype data) by testing whether the recombination rate between the two loci is significantly different from 1/2 ($\theta \neq \frac{1}{2}$). The advantage of considering

short pedigrees is that only a few recombinations can occur between loci in linkage so that sparse marker information (up to hundreds of markers per chromosome) are sufficient to map disease variants, although with limited resolution. Examples for some of the first linkage analyses include Alzheimer's disease (Schellenberg et al., 1991), cystic fibrosis (Lathrop et al., 1988) and Huntington's disease (Gusella, 1984). However, the method has been less successful in mapping genes for complex disorders (Altmüller et al., 2001).

1.2.3 Genome-wide association studies

High-throughput genotyping technologies, including SNP arrays and low-coverage sequencing, have enabled genotyping of hundred of thousands to millions of common SNPs in increasing sample sizes. These technological advances, together with the findings from the Hapmap project on the haploblock structure of the human genome (see Section 1.1.3) laid the foundation for the success of genome-wide association studies (GWAS, Frazer et al., 2009), currently the most widely used design for gene mapping (McCarthy et al., 2008). GWAS rely on the LD between genotyped and causal variants, which are potentially not typed, to identify loci implicated in traits and diseases. Contrarily to linkage studies, GWAS are commonly performed in populations of unrelated individuals (Astle and Balding, 2009), enabling the analysis of larger sample sizes (typically from 1,000 to 100,000 individuals or even larger[2]). The basic principle underlying GWAS is to test for association between individual genotyped variants and the trait of interest. Given the large number of genome-wide variants to be tested, stringent criteria of significance are needed to control the number of false discoveries at the expense of statistical power (see Section 2.2.1).

The first GWAS in human was published in 2005 and revealed two genetic loci associated with age-related macular degeneration (Klein et al., 2005). Ever since, GWAS have become increasingly popular, yielding important insights into the genetic architecture of many complex traits, including type 2 diabetes (Scott et al., 2007), inflammatory bowel disease (Khor et al., 2011) and major depression (Kohli et al., 2011; Cai et al., 2015). Today, the GWAS Catalog (Burdett et al., 2015) includes more than 2,500 published GWA studies and more than 24,000 SNP-trait associations (August 2016).

Importantly, dense genotype data, such as those from the HapMap, the 1000 Genomes project and, more recently, the UK10K population (Huang et al., 2015) can

[2]Consortia such as UK Biobank (Sudlow et al., 2015) have currently been producing and studying cohorts of 500,000 individuals.

be used as a reference for imputing genotypes at unobserved loci, thereby completing sparse genotype data (Howie et al., 2009; Marchini and Howie, 2010). Recent GWA studies in cohorts with NGS genotype data have attempted to identify effects from rare variants (UK10K Consortium, 2015). However, the robust identification of rare-variant effects remains challenging, as it requires extremely large sample sizes (Risch and Merikangas, 1996; Bush and Moore, 2012).

1.2.4 Molecular mechanisms from genotype to phenotype

In the process of gene expression, the genetic information stored in the DNA is used to produce molecular products. The information on the composition of these products is contained in restricted portions of the genome known as genes. In the first step of gene-expression (transcription), genes are transcribed into ribonucleic acid (RNA) molecules. Although RNA molecules can be the final product of gene expression, many RNA molecules are translated into proteins in a process known as translation. Proteins are long chains of smaller molecules called amino acids. Only approximately 1.5% of the human genome codes for proteins (Lander et al., 2001) while a large fraction of the remaining portion is likely to play a role in the regulation of gene expression (ENCODE Project Consortium, 2004). The impact of genetic variation on phenotypes is the consequence of perturbations to this complex molecular machinery.

An easily interpretable mechanism through which genetic variants may affect phenotype is the direct alteration of the structure of the coded protein and thus of its functionality. For example, *sickle cell anaemia* is caused by a SNP in the *HBB* gene, which causes a substitution of an amino acid in the sequence of the coded protein (Laird and Lange, 2010). Alternatively, genetic variation may affect the regulation of gene expression. One way this can occur is through the disruption of a specific sequence that affects the binding of proteins regulating the expression of a gene. Another possibility is the alteration of the structure of the DNA, thereby affecting the functionality of regulatory elements and ultimately gene expression (ENCODE Project Consortium, 2004; Kundaje et al., 2015). Importantly, these structural changes are not necessarily associated with a change in the DNA sequence and are collectively referred to as epigenetics, where the prefix *epi-* is from Greek and means "outside of" (Spector, 2012). Importantly, such changes can be "inherited" in the process of cellular division and play a key role in regulation of gene expression. Examples of epigenetic modifications are DNA methylation and histone modifications. Specifically, DNA methylation is the process through which a hydrogen atom in adenine or cytosine is replaced by a methyl group. Methylation of promoters (i.e., the DNA sequences where transcription is initi-

ated) is associated with the silencing of the corresponding gene (Alberts et al., 2014). Another important class of epigenetic changes are modifications of structural proteins, known as histones, which play a key role in the packing of DNA in chromosomes. Notably, histone proteins can undergo more than 100 histone modifications (Ernst and Kellis, 2010), which can be reversibly written and erased by specific enzymes. The combination of the different histone modifications determines "the histone code" (Alberts et al., 2014). Different combinations have been associated with transcribed regions, enhancers, promoters, and other functional regions (Ernst and Kellis, 2010). Regulatory elements can lie far from the regulated gene (Alberts et al., 2014), hampering the identification of genes underlying known GWAS loci.

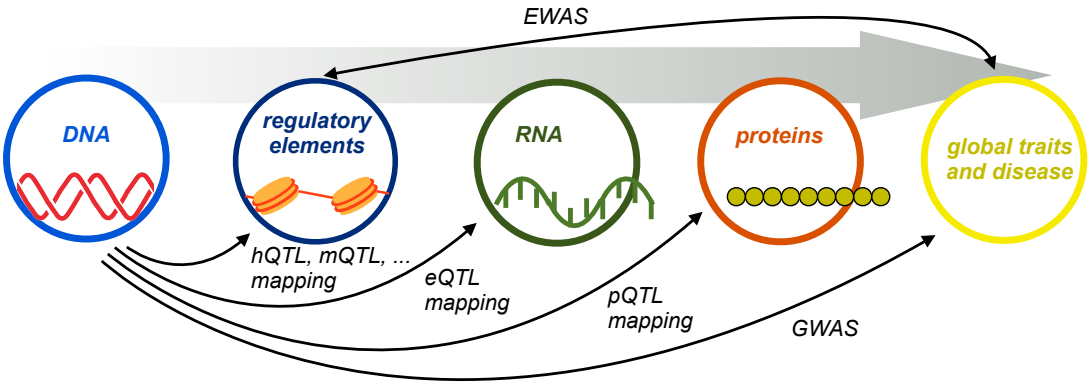


Figure 1.2: **Overview of genetic analysis of molecular and global traits.** Shown are different molecular layers and association analyses that can be considered to link genetic variation to different molecular and global traits. The grey arrow shows the most likely causal direction of these effects. Of course there are numerous exceptions to this simplified scheme. Regulatory elements include DNA methylation, histone modification, binding affinity for different proteins, etc. hQTL, mQTL, eQTL, pQTS mapping stand for histone modification, methylation, expression and protein QTL mapping. In order to map genes whose regulation is associated with phenotypic changes, many studies have considered genome-wide association testing between epigenetic features and complex traits, including disease, mainly focusing on DNA methylation (Paul and Beck, 2014). These studies are known as epigenome-wide association studies (EWAS).

The decrease in cost of high-throughput profiling of gene expression has made it possible to measure gene expression levels in large numbers of individuals, thereby enabling the mapping of quantitative trait loci for gene expression (eQTL mapping, Schadt et al., 2003). More recently, these studies have been extended to test for genetic effects on different regulatory elements, including DNA methylation (Gaunt et al., 2016), histone modifications (Grubert et al., 2015) as well as the activity of known

proteins and protein complexes that regulate expression and chromatin accessibility, such as transcription factors (Waszak et al., 2015). An overview of the different genetic analyses and their relationship to GWAS is given in **Fig. 1.2**.

As genetic effects on molecular traits may depend on factors such as tissue, environment and cell type, it is necessary to measure expression and other molecular traits in disparate cellular contexts. For this reason, the Genotype-Tissue Expression Project (GTEx Consortium, 2015) has been collecting and analysing gene expression profiles for more than 50 tissue types from over 900 donors.

Molecular QTL mapping entails QTL mapping for tens of thousands of molecular measurements. Therefore the "small sample size/high number of tests" problem in these analyses is even more severe than in GWAS. Moreover, genetic analysis of molecular traits are typically performed in datasets with smaller sample sizes compared to GWAS[3]. To circumvent this issue, many studies have focused on mapping proximal (putatively *cis*-acting) QTLs by restricting association testing to genetic variants that are close to the analysed molecular trait as they are more likely to have an effect on the molecular phenotype compared to distal genetic variants.

1.3 Statistical challenges in genetics

Despite ever increasing sample sizes, the statistical analysis of genetic data remains challenging. In this section I discuss the limitations of the standard GWAS approach and recent extensions focusing on aspects that are relevant to the contributions of this thesis.

Polygenic architectures. A main limitation of the GWAS methodology is that many complex traits have highly polygenic architectures with numerous weak-effect variants (Frazer et al., 2009). This is one potential source of "missing heritability" (Maher, 2008), where heritability is defined as the fraction of phenotypic variance explained by additive genetic effects. This narrow-sense definition of heritability does not include dominance effects and genetic interactions, which are included in the definition of broad-sense heritability (Visscher et al., 2008). Heritability was traditionally estimated from phenotype data on pedigrees (see Falconer and Mackay, 1996), for example by regressing the parental average trait against the trait in the offspring (Laird and Lange, 2010). The phenomenon of "missing heritability" refers to the positive difference between these heritability estimates and the proportion of phenotypic variance explained

[3]Typically, molecular QTL mapping is performed considering a few hundreds of individuals.

by the additive effects of GWAS loci. Performing GWAS in increasingly large cohorts has helped recover some of this missing heritability, by revealing new associated loci for traits such as height (Wood et al., 2014), intelligence (Rietveld et al., 2013) and schizophrenia (Ripke et al., 2014). An alternative approach is to jointly model the effects from multiple genetic variants. This strategy has been shown to recover large proportions of the missing heritability for many complex traits (Yang et al., 2010; Lee et al., 2012b; Gusev et al., 2014; Eichler et al., 2010). A widely used approach to aggregate genetic effects across multiple variants is to use a random effect within a linear mixed model (LMM), the same approach initially proposed by Fisher to model inheritance of complex traits (Fisher, 1918). Prior to applications in human studies and following Fisher and Henderson's work, LMMs had been extensively used in the field of animal breeding (Fisher, 1921; Fisher and Mackenzie, 1923; Henderson, 1984).

Recent studies have shown that genetic effects are not uniformly distributed along the genome (Gusev et al., 2013) and that genetic loci can harbour multiple causal variants (Wood et al., 2011; Chiba-Falek et al., 2012; Trynka et al., 2011; Ehret et al., 2012; Patsopoulos et al., 2013; Corradin et al., 2014). These findings have motivated the development of set tests, a class of models that allows joint testing of multiple variants in genetic regions (Wu et al., 2011; Listgarten et al., 2013; Brown et al., 2016). Set tests have been shown to improve genetic mapping over single-variant approaches, recovering parts of the unexplained genetic variance (Wu et al., 2010; Listgarten et al., 2013; Brown et al., 2016).

Correcting for confounding. A second challenge in GWAS is confounding factors, which can lead to spurious genotype-trait associations (McClellan and King, 2010; Lambert and Black, 2012; Pritchard et al., 2000b; Patterson et al., 2006). A well-studied confounder in association studies is population structure (Marchini et al., 2004). To showcase its action, let us suppose we have a population dataset consisting of individuals from different ethnic groups. As alleles at many different loci tend to co-occur in individuals from the same ethnic group, the genotype data will be characterised by genome-wide LD between variants. As shown in **Fig. 1.3a**, if we consider a quantitative trait that is also influenced by ethnicity (for example because it is affected by an environmental factor determined by the geographic region of origin), a GWAS will retrieve many spurious genetic associations (see Lander and Schork, 1994). In this example, population structure exhibits a shared influence on both phenotype and genotype data. As shown in **Fig. 1.3b**, population structure can also cause genuine genetic signal from causal variants to be mirrored in numerous non-causal loci in LD (Ewens and Spielman,

1995). Similar problems affect genetic analyses of related individuals, where genome-wide genetic similarities are correlated with environmental factors, thereby creating many spurious genotype-trait associations (Eu-Ahsunthornwattana et al., 2014).

As discussed in further detail in the next chapter, several methods have been proposed to correct for such confounding factors. Among these different strategies, the linear mixed model has emerged as a particularly robust approach as it can handle different types of confounding, including complex population structure and relatedness (Kang et al., 2008b; Kang et al., 2010). However, fitting LMMs can be generally computationally demanding. Although recent computational advances have enabled application of specific LMMs to genetic analyses in large cohorts (Kang et al., 2008b; Zhou and Stephens, 2012; Kang et al., 2010; Lippert et al., 2014a), the application of more complex LMMs is challenging due to a large computational burden.

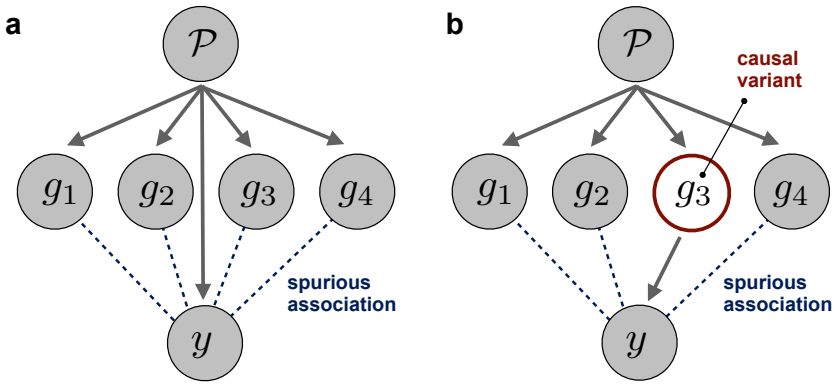


Figure 1.3: **Two causal networks of confounding.** Population structure ($\mathcal{P}$) affects genome-wide genotypes, creating long-range LD between variants. In (**a**) $\mathcal{P}$ also affects the trait of interest (y), a scenario that creates spurious associations in GWAS. In (**b**) $\mathcal{P}$ does not have an effect on the trait, however, because of long-range LD, the effect of the causal genotype g_3 can be mirrored at physically distant loci in an association test.

Multivariate analysis of related phenotypes. As increasingly deep phenotype data are becoming available in human cohorts, there has been a growing interest in multivariate analyses. Multivariate models have several benefits. First, many complex traits and diseases are affected by shared genetic and environmental influences (Fortune et al., 2015; Frazer et al., 2009). For example, Pickrell et al. (2016) found that a notable proportion of the variants affecting age at menarche also affect height (36%), age of voice drop (30%), BMI (28%), breast size (10%) and male pattern baldness (10%).

The phenomenon of a variant affecting multiple traits is also known as pleiotropy and its occurrence in human traits is pervasive (Sivakumaran et al., 2011). Joint genetic analyses of multiple traits have been shown to increase statistical power by leveraging trait-to-trait correlations induced by shared genetic and environmental factors (Korte et al., 2012; Zhou and Stephens, 2014). Second, genetic effects on phenotypes can depend on environment, sex, age and other contextual variables (Dick, 2011; Winkler et al., 2015; Sasaki et al., 2015; Kang et al., 2014). By modelling the same trait in varying contexts, multivariate approaches can help characterising the context-specific architecture of genetic effects (Korte et al., 2012; Kang et al., 2014). Beyond the analysis of global phenotypes, these multivariate approaches are also important to characterise the genetic regulation of molecular traits across tissues (Price et al., 2011; Sul et al., 2013), environments (Smith and Kruglyak, 2008), development (Francesconi and Lehner, 2014) and external stimuli (Fairfax et al., 2014). Finally, multivariate analyses can be used to investigate the molecular mechanism of genetic effects by enabling joint analyses across multiple phenotypic and molecular layers (Wallace et al., 2012; Pickrell et al., 2016; Giambartolomei et al., 2014). Despite these benefits, many multivariate models are hampered by poor computational scalability, hindering their application to larger cohorts. For example, variant-set tests across multiple traits were limited to analyses of small cohorts prior to the work presented in this thesis. Additionally, the number of possible integrative analyses grows dramatically with the dimensionality of the phenotypic data.

1.4 Thesis overview and individual contributions

The challenges described in Section 1.3 illustrate the limitations of the "one-variant-to-one-trait" approach of standard GWAS and highlight the need for novel integrative methods. The linear mixed model has emerged as a flexible framework for many genetic analyses. Linear mixed models ensure robust control of confounding effects and have been widely used for association testing for single-variants (Kang et al., 2008b; Zhou and Stephens, 2012; Sikorska et al., 2013) and variant-sets (Listgarten et al., 2013; Lippert et al., 2014a). Moreover, recent computational advances have enabled efficient single-variant association tests across multiple traits (Zhou and Stephens, 2014; Lippert et al., 2014b; Furlotte and Eskin, 2015). Despite these developments, two main factors have limited application of a larger class of LMMs to genetic analyses. First, with the exception of some particular cases, LMMs are generally computationally inefficient. Second, available software are designed to perform very specific analyses while frame-

works that enable the flexible design of genetic models are not widely available (an exception is Gilmour et al., 2009). In this thesis, I present novel and efficient LMMs that can model relationships between multiple variants and traits, and describe an inference framework that enables LMMs to be built flexibly.

In Chapter 2, I give an overview of current LMMs for genetic analyses, covering their use for association testing, heritability estimation, variance decomposition, set tests and multi-trait modelling.

In Chapter 3 I present an extension of existing LMM inference schemes that enables association testing between multiple variants and traits in large cohorts (mtSet). After demonstrating and benchmarking the model in extensive simulations, I discuss two applications to real data. This work was done in collaboration with Barbara Rakitsch, Christoph Lippert and Oliver Stegle and resulted in the following publication

- Francesco Paolo Casale*, Barbara Rakitsch*, Christoph Lippert, and Oliver Stegle. "Efficient set tests for the genetic analysis of correlated traits." Nature methods 12, no. 8 (2015): 755-758.
  * Joint first authorship

  Individual contributions:
  Francesco Paolo Casale and Barbara Rakitsch developed the method. Francesco Paolo Casale and Barbara Rakitsch analysed the data. Christoph Lippert provided analysis tools and contributed to the interpretation of results.

Building on the mtSet framework, in Chapter 4, I derive novel tests for interactions between variant-sets and categorical contexts (iSet). iSet can be used for interaction testing either in (i) datasets where trait measurements are available in the same set of individuals in different contexts or (ii) by stratifying a population into distinct subgroups based on an external context variable. In extensive simulations, I demonstrate that iSet offers power advantages compared to previous interaction tests. Additionally, the model can identify regions associated with changes in the configuration of causal variants across the analysed contexts. I discuss results from applications of iSet to a monocyte stimulus eQTL study (Fairfax et al., 2014) and a gene-by-sex interaction analysis of blood lipid traits. The work was done in collaboration with Danilo Horta, Barbara Rakitsch and Oliver Stegle and resulted in the following publication

- Francesco Paolo Casale, Danilo Horta, Barbara Rakitsch, and Oliver Stegle.

"Joint genetic analysis using variant sets reveals polygenic gene-context interactions." PLoS genetics 13.4 (2017): e1006693.

Individual contributions:
Francesco Paolo Casale developed the method and analysed the data. Barbara Rakitsch and Danilo Horta provided analysis tools.

In Chapter 5 I present LIMIX, a mixed-model framework that allows performing different genetic analyses in one tool. LIMIX has been used in several projects to conduct variance component analyses (Dubin et al., 2015a; Sasaki et al., 2015; Baud et al., 2017; Chen et al., 2016), single- and multi-trait association test (Kawakatsu et al., 2016; Horton et al., 2016; Sudmant et al., 2015; Schor et al., 2017; Cannavò et al., 2016) and genomic predictions (Märtens et al., 2016). The advantage of LIMIX over other tools is its flexibility, which allows designing customised models to best suit the scope and data of specific studies. All the models considered in this thesis are implemented within the LIMIX software framework, including both new methods and a wide range of existing approaches that were used for comparison. Finally, to showcase the importance of flexible modelling to investigate high dimensional data I present two applied analyses from collaborative projects. The first analysis is part of a collaborative project with Jacob Degner, Ignacio Shor, Oliver Stegle Ewan Birney and Eileen Furlong's group at EMBL in Heidelberg, Germany (see Schor et al., 2017). The aim of this project is to understand the impact of genetic variation on transcription initiation in *Drosophila melanogaster* during development. The second analysis is a component of the BluePrint-WP10 project (Chen et al., 2016) and done in collaboration with Lu Chen, Oliver Stegle, Nicole Soranzo and others. In this analysis, we leverage high-resolution genetic, epigenetic and transcriptomic data in three human immune cell types (I will show results only from one cell type in this thesis) to quantify the contribution of *cis*-genetic and *cis*-epigenetic effects to gene expression variability.

- LIMIX framework.
  Individual contributions:
  Francesco Paolo Casale, Danilo Horta and Barbara Rakitsch wrote the source code for flexible inference. Francesco Paolo Casale wrote the source code for the set tests and the variance decomposition module. Christoph Lippert and Oliver Stegle wrote the source code for fixed effect testing.

- QTL mapping of transcription initiation in Drosophila development.
  Individual contributions:

Francesco Paolo Casale, Jacob Degner and Oliver Stegle designed the statistical models. Jacob Degner and Francesco Paolo Casale performed the QTL mapping. Igacio Shor, Jacob Degner, Eileen Furlong and Oliver Stegle interpreted the results.

- Dissecting genetics and epigenetics effects in three immune cell types.
  Individual contributions:
  Francesco Paolo Casale, Oliver Stegle and Nicole Soranzo designed the statistical models. Francesco Paolo Casale performed the analysis. Nicole Soranzo, Oliver Stegle and Francesco Paolo Casale interpreted the results.

Finally, in Chapter 6, I give a summary of the work in this thesis and provide an outlook on future research.

A full list of publications can be found in Section E.

2 | Linear mixed models for genetic analyses

The linear mixed model (LMM) has become the standard framework for many genetic analyses. LMMs provide robust control for confounding factors, allow for aggregating genetic effects from multiple variants and enable the joint analysis of multiple traits. While inference in LMMs is in general computationally demanding, efficient implementations of specific LMMs enable applications to large datasets. In this chapter, I give an overview of the use of LMMs in genetics and efficient algorithmic implementations. In Sections 2.1-2.2, I discuss linear models and basic concepts of genome-wide association studies (GWAS). In Section 2.3, I introduce the LMM and discuss applications in genetics. Finally, in Section 2.4, I present the extension of LMMs to the analysis of multiple traits.

2.1 Linear regression

A linear model describes a continuous output variable as a linear function of one or more input variables (also referred to as features). Denoting with N the number of samples, y_i the output variable for sample i and $\{x_{i1}, \ldots, x_{iF}\}$ F input variables for sample i, the linear model can be cast as

$$y_i = \sum_{f=1}^{F} x_{if}\beta_f + \psi_i, \quad \text{with } \psi_i \sim \mathcal{N}\left(0\,,\,\sigma_e^2\right). \tag{2.1}$$

The residual term ψ_i accounts for the fact that the x-y relationship is not deterministic because of measurement noise or other unmodelled factors. ψ_i is here assumed to follow a normal distribution with mean 0 and variance σ_e^2 and to be independent across samples, i.e. $\text{cov}(\psi_i, \psi_j) = 0$. In equation (2.1), β_f denotes the weight of the input feature f.

Introducing the output vector $\boldsymbol{y}$, the input matrix $\boldsymbol{X}$, the weight vector $\boldsymbol{\beta}$ and the residual vector $\boldsymbol{\psi}$ as

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1F} \\ x_{21} & x_{22} & \dots & x_{2F} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \dots & \dots & x_{NF} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_F \end{bmatrix} \quad \text{and} \quad \boldsymbol{\psi} = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_N \end{bmatrix}, \tag{2.2}$$

the linear model in (2.1) can be expressed in matrix form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\psi}, \quad \text{with } \boldsymbol{\psi} \sim \mathcal{N}\left(\boldsymbol{0}\,,\, \sigma_e^2 \boldsymbol{I}_N\right), \tag{2.3}$$

where $\boldsymbol{I}_N$ denotes the $N \times N$ identity matrix.

2.1.1 Maximum likelihood solution

Equation (2.3) specifies the probability distribution of the data $p\left(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{\beta}, \sigma_e^2\right)$ given the input variables $\boldsymbol{X}$ and the model parameters $\boldsymbol{\beta}$ and σ_e^2. This probability is known as the likelihood of the model and, for parameter inference, is typically regarded as a function of the model parameters and denoted as $\mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right)$. The model in (2.3) can thus be equivalently specified as

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right) = p\left(\boldsymbol{y} \,|\, \boldsymbol{X}, \boldsymbol{\beta}, \sigma_e^2\right) = \mathcal{N}\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}\,,\, \sigma_e^2 \boldsymbol{I}_N\right), \tag{2.4}$$

or more directly as

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta}\,,\, \sigma_e^2 \boldsymbol{I}_N\right). \tag{2.5}$$

The log marginal likelihood of the model can be explicitly expressed as

$$\log \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right) = -\frac{N}{2} \log\left(2\pi\right) - \frac{1}{2} N \log \sigma_e^2 - \frac{1}{2\sigma_e^2} \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right). \tag{2.6}$$

The maximum likelihood estimator (MLE) of the model parameters is defined as the set of parameter values that maximise the likelihood (or its log). Denoting with $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_e^2$ the MLE of $\boldsymbol{\beta}$ and σ_e^2 we can write

$$\hat{\boldsymbol{\beta}}, \hat{\sigma}_e^2 = \text{argmax}_{\boldsymbol{\beta}, \sigma_e^2}\, \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right). \tag{2.7}$$

The MLE of $\boldsymbol{\beta}$ and σ_e^2 can be found by equating the gradients of the log likelihood to 0

$$\left(\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right)}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \sigma_e^2=\hat{\sigma}_e^2} = 0 \tag{2.8}$$

$$\left(\frac{\partial \log \mathcal{L}\left(\boldsymbol{\beta}, \sigma_e^2\right)}{\partial \sigma_e^2}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \sigma_e^2=\hat{\sigma}_e^2} = 0 \tag{2.9}$$

From these two equations it can be easily shown that

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{2.10}$$

$$\hat{\sigma}_e^2 = \frac{1}{N}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \tag{2.11}$$

$$= \frac{1}{N}\left(\boldsymbol{y}-\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right)^\top \left(\boldsymbol{y}-\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right) \tag{2.12}$$

2.1.1.1 Restricted maximum likelihood

The MLE of variance parameters is biased in Gaussian models as consequence of the fact that the weights are estimated from the data, which entails a reduction of the effective number of degrees of freedom. Patterson and Thompson (1971) proposed to estimate variance parameters by maximising the restricted (or residual) maximum likelihood (REML), which can be obtained by projecting the output vector in a space that is orthogonal to $\boldsymbol{X}$. Considering Eq (A.3) for the model in Eq (2.5), we obtain the following log restricted maximum likelihood (see Section A.1)

$$\log \mathcal{L}\left(\sigma_e^2\right) = -\frac{N-F}{2}\log\left(2\pi\right) - \frac{1}{2}\log\det\left(\boldsymbol{X}^\top \boldsymbol{X}\right) \tag{2.13}$$

$$-\frac{1}{2}(N-F)\log\sigma_e^2 - \frac{1}{2\sigma_e^2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \tag{2.14}$$

that is maximised by

$$\hat{\sigma}_e^{(\mathrm{REML})\,2} = \frac{1}{N-F}\left(\boldsymbol{y}-\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right)^\top \left(\boldsymbol{y}-\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\right). \tag{2.15}$$

Eq (2.15) is identical to Eq (2.12) with the exception that N is replaced by $N-F$, which denotes the loss of F degrees of freedom.

2.2 Linear models for genome-wide association studies

In genetics, the output variable is typically the trait of interest while input variables can include genetic variants and known factors, such as age and sex, which can have an influence on the trait. The linear model that has been widely used in GWAS, models the phenotype vector as the sum of the contributions of the variant being tested, the contribution from K known factors (covariates), and residual noise

$$\boldsymbol{y} = \underbrace{\boldsymbol{g}\beta}_{\text{variant effect}} + \underbrace{\boldsymbol{X}\boldsymbol{\alpha}}_{\text{covariate effects}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}}, \quad \boldsymbol{\psi} \sim \mathcal{N}(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I}_N), \tag{2.16}$$

where I explicitly separated the terms corresponding to the genetic variant and the covariates. In equation (2.16), $\boldsymbol{y} \in \mathbb{R}^{N\times 1}$ denotes the phenotype vector for N individuals, $\boldsymbol{g} \in \mathbb{R}^{N\times 1}$ denotes the genotype vector of the tested variant, $\boldsymbol{X} \in \mathbb{R}^{N\times K}$ denotes the input matrix for K covariates, $\beta \in \mathbb{R}$ denotes the weight of the variant (also referred to as the effect size of the variant), and $\boldsymbol{\alpha} \in \mathbb{R}^K$ denotes the weight vector of the covariates. Note that for a diploid organism, the representation of genotypes as numerical values requires making some assumptions on the genetic model. Let us consider a bi-allelic variant with major allele a and minor allele A. For the minor allele A, we can consider either a dominant model ($aa = 0$, $Aa = 1$, $AA = 1$; where only one copy of the allele is necessary to have a phenotypic effect), a recessive model, ($aa = 0$, $Aa = 0$, $AA = 1$; where two copies of the minor allele must be present for a phenotypic effect) or an additive model ($aa = 0$, $Aa = 1$, $AA = 2$; where the effect is proportional to the minor allele count). In this thesis, I will consider an additive genetic model, which is widely-used in the analysis of complex traits (Laird and Lange, 2010).

2.2.1 Statistical hypothesis testing

Association testing between a trait and a genetic variant can be assessed by comparing the hypothesis that the variant has an effect, $\beta \neq 0$, on the trait ($\mathcal{H}_1$) versus the hypothesis that the variant has no effect ($\mathcal{H}_0$)

$$\mathcal{H}_1 : \quad \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{g}\beta + \boldsymbol{X}\boldsymbol{\alpha}\,,\, \sigma_e^2 \boldsymbol{I}_N\right), \tag{2.17}$$

$$\mathcal{H}_0 : \quad \boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha}\,,\, \sigma_e^2 \boldsymbol{I}_N\right). \tag{2.18}$$

$\mathcal{H}_1$ and $\mathcal{H}_0$ are referred to as the alternative and the null hypothesis of the statistical test, respectively. Statistical hypothesis testing consists of three basic steps. First, we calculate a test statistic, which is a random variable that quantifies the evidence that

$\mathcal{H}_1$ is true. Second, we calculate a probability value (P value) as the probability, under $\mathcal{H}_0$, of sampling a test statistic at least as extreme as the observed one. The P value is a function of the test statistic and, by definition, it is uniformly distributed under $\mathcal{H}_0$. Finally, if this probability is low $\mathcal{H}_0$ is rejected and $\mathcal{H}_1$ accepted (positive result), otherwise, we reject $\mathcal{H}_1$ and accept $\mathcal{H}_0$ (negative result). In statistical hypothesis testing, two types of errors can be made. We can either reject $\mathcal{H}_0$ when $\mathcal{H}_0$ is true, thus generating a false positive (type-I error), or reject $\mathcal{H}_1$ when $\mathcal{H}_1$ is true, generating a false negative (type-II error). Other concepts that are central to statistical hypothesis testing are the significance level, defined as the type-I error rate (i.e. the expected percentage of false positives), and the statistical power, which is the true positive rate under $\mathcal{H}_1$ (i.e. the ability to recover true associations).

A commonly used test statistic is the log likelihood ratio (LLR), which we will consider throughout this thesis. The LLR test statistic D is defined as

$$D = \log \mathcal{L}\left(\hat{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}_e^2\right) - \log \mathcal{L}\left(0, \bar{\boldsymbol{\alpha}}, \bar{\sigma}_e^2\right), \tag{2.19}$$

where $\log \mathcal{L}\left(\hat{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}_e^2\right)$ is the log-likelihood of the alternative model, $\left\{\hat{\beta}, \hat{\boldsymbol{\alpha}}, \hat{\sigma}_e^2\right\}$ the MLE of the parameters under the alternative model and $\left\{\bar{\boldsymbol{\alpha}}, \bar{\sigma}_e^2\right\}$ the MLE of the parameters under the null model[1]. A convenient theorem from Samuel Wilks (Wilks, 1938) guarantees that under asymptotic assumptions (i.e. infinite sample size) and when null model parameters are not at the bound of the domain of the likelihood of the alternative model, $2D$ follows a χ^2 distribution with number of degrees of freedom d equal to the number of tested parameters ($2D \sim \chi^2(d)$). The P value, for a d degree-of-freedom (dof) test, can thus be calculated from the observed LLR test statistic D as

$$P(D) = \int_{2D}^{\infty} \chi^2\left(x;\ d\right) dx = 1 - F_{\chi^2}(2D;\ d), \tag{2.20}$$

where F_{χ^2} is the cumulative density function of the χ^2 distribution. A single-variant test ($\beta \neq 0$) has one degree of freedom, $d = 1$.

2.2.2 Multiple hypothesis testing correction

Hundreds of thousands or millions of variants may be individually tested within a typical human GWAS. When performing such a large number of tests, controlling single-test P values results in a high number of false positives (for example, for $P < 0.01$ and 10^6 tests we expect $10,000$ false positives under the null hypothesis). This problem

[1]Note that the log-likelihood function of the null model is $\log \mathcal{L}\left(\beta = 0, \boldsymbol{\alpha}, \sigma_e^2\right)$.

is known as the multiple hypothesis testing problem. In the following, I give a brief overview of the methods commonly used in genetic analysis to correct for multiple hypothesis testing.

Controlling family-wise error rate. One strategy is to control the probability of having at least one false positive in the experiment, which corresponds to an experiment-wise P value known as family-wise error rate (FWER)[2].

The widely used Bonferroni method follows this strategy assuming independence between tests. Given a desired family-wise significance level $\bar{\alpha}$, the method consists in calculating adjusted P values $\bar{P} = Pn$, where n is the number of tests, and setting $\bar{P} < \bar{\alpha}$. This strategy ensures $FWER < \bar{\alpha}$. The Bonferroni correction strategy is conservative, as the consequence of the assumption of independence between test, which ignores correlations between genotypes due to linkage disequilibrium (LD). An alternative strategy, which accounts for the dependency of the statistical tests, is to consider permutations. For example, one way to control the FWER by using permutations is to perform the experiment M times, each time considering a different permutation of the genotype data across individuals. The minimum P values from these M additional experiments are then used to calculate an experiment-wise P value, as the fraction of the M minimum permutation P values that are lower than the minimum observed P value. This approach has been used in *cis* molecular QTL mapping to estimate gene-level P values (Sudmant et al., 2015; GTEx Consortium, 2015). Although this strategy accounts for local LD, thereby increasing the statistical power, it entails a great computational burden and can become unpractical in molecular analyses of large cohorts.

Recently, several permutation-free methods that allow accounting for local LD have been proposed (Xu et al., 2014; Sul et al., 2015; Davis et al., 2016). Davis et al. (2016) proposed to estimate the effective number of independent tests from the genotype data. Specifically, denoting with R the number of variants in the considered genetic region and with $\boldsymbol{G}$ the $N \times R$ genotype matrix (encoded as minor allele counts), Davis et al. (2016) consider a regularised estimator $\hat{\boldsymbol{\Sigma}}$ of the correlation matrix between the columns of $\boldsymbol{G}$. This estimator was initially proposed by Ledoit and Wolf (2004) and has been shown to produce well-conditioned matrices in cases where $R > N$. The number of effective tests is then estimated as the minimum number of eigenvalues of $\hat{\boldsymbol{\Sigma}}$ needed to explain a certain fraction C of the total variance (the suggested value for C is 99%). This method is called eigenMT and I will use it in Chapter 4.

[2]Denoting with n the number of tests and with α the type I error rate we have FWER $= 1-(1-\alpha)^n$.

Controlling the false discovery rate. An alternative solution is to control the false discovery rate (FDR), i.e. the expected percentage of false discoveries.

The most widely used FDR-based correction method is the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), which again assumes independence between tests[3]. Let us consider T tests with P values $p_1, p_2, \ldots, p_T$ and let $\nu_1, \nu_2, \ldots, \nu_T$ be their ranks (the smallest P value has rank 1, the highest has rank T), defining adjusted P values as $\bar{p}_i = \frac{Tp_i}{\nu_i}$ and setting $\bar{p}_i < \alpha$ ensures FDR $< \alpha$.

Multiple hypothesis testing correction in molecular *cis*-QTL mapping. A typical strategy to correct for multiple hypothesis testing in molecular *cis*-QTL mapping is to use a two-step procedure (Battle et al., 2014; Sudmant et al., 2015; GTEx Consortium, 2015). First, for each gene an experiment-wise P value is obtained by correcting for multiple testing across variants using a FWER-based method. These gene-level P values are probability values for the hypothesis of a gene having at least one QTL in the analysed region. Second, the gene-level P values are corrected to control the FDR, typically using the Benjamini-Hochberg procedure.

2.2.3 Distribution of P values and QQ plot

Under the assumption that the vast majority of the tested variants are not associated with the analysed trait, a GWAS is expected to produce approximately uniform P values. A representation that is typically used to compare the observed and the expected distributions of P values is the quantile-quantile plot (QQ plot). In a QQ plot the observed $-\log_{10}P$ is plotted against the expected $-\log_{10}P$, where the expected value is obtained from the uniform distribution[4]. **Fig. 2.1** shows a close-to-ideal P value distribution and the corresponding QQ plot from a GWAS of simulated data. In this example, only a few variants are significantly associated with the trait and deviate from the uniform distribution. A quantitative measure of the discrepancy between the observed and the expected distributions is the genomic control λ_{GC} (Devlin and Roeder, 1999), also known as genomic inflation factor. A definition of genomic control is $\lambda_{\text{GC}} = \text{median}\left(\log_{10}(P)\right) \ / \ \log_{10}\left(0.5\right)$, i.e. it is the the ratio of the median of the expected and the observed distributions of the log P values. Inflated QQ plots (anti-conservative P values) are associate to $\lambda_{\text{GC}} > 1$ while deflated QQ plots (conservative P values) correspond to $\lambda_{\text{GC}} < 1$. As confounding factors such as population

[3]However, the Benjamini-Hochberg procedure is still valid under different dependence assumptions (Sun and Tony Cai, 2009).

[4]Let $p_1 \leq \cdots \leq p_T$ be the P values from T tests the expected P values for p_i is $p_i^{(\text{exp})} = \frac{i}{T+1}$ if all tests are under the null hypothesis.

structure and relatedness create spurious genome-wide associations, inflated QQ plots are typically associated with the presence of confounding (Voight and Pritchard, 2005; Lin and Sullivan, 2009). However, in analyses of highly polygenic traits, inflation can arise from genuine genetic signal (Yang et al., 2011c). Bulik-Sullivan et al. (2015) have recently proposed a method to quantify the inflation that can be attributed to confounding only. The discrimination between polygenicity and confounding is based on the intuition that only polygenicity is associated with high-LD regions.

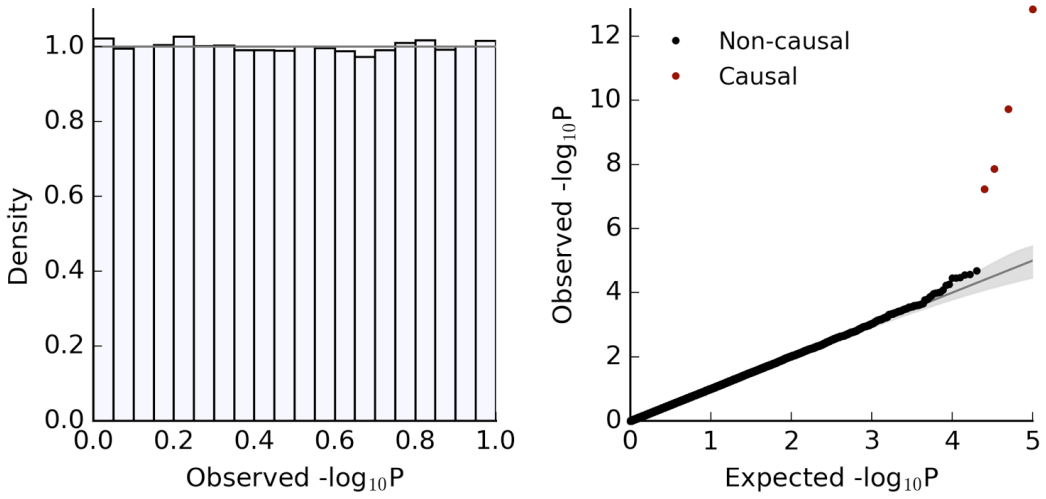


Figure 2.1: **Example of P value distribution and QQ plot for a simulated GWAS.** I simulated genotype data for $N = 500$ individuals and $S = 100,000$ variants as $x_{ij} \sim \text{Bernoulli}\,(n_t, r)$ with number of trials $n_t = 2$ and rate $r = 0.2$, randomly selected $S_c = 4$ variants as causal and generated the phenotype as $\boldsymbol{y} = \sum_{j=1}^{S_c} \boldsymbol{x}_j \beta_j + \boldsymbol{\psi}$ where $\boldsymbol{x}_j$ is the standardised genotype vector of causal variant j, $\beta_j = \pm\sqrt{0.1}$ and $\psi \sim \mathcal{N}\left(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N\right)$, with $\sigma_e^2 = 0.6$. The left panel shows the distribution of the association P values obtained for the S variants while the right panel shows the corresponding QQ plot (the 4 causal variants are highlighted in red). The shaded area indicates the 99% confidence interval around the diagonal (dark grey line).

2.2.4 Accounting for confounding in the linear model

The first two methods to account for population structure were genomic control (Devlin and Roeder, 1999) and structured association (Pritchard et al., 2000a). Genomic control correction adjusts for inflation by dividing the test statistic of each marker by the genomic control parameter. However, as different markers have different abilities to differentiate across populations, this uniform adjustment is far from optimal. Indeed,

the test statistic of markers that strongly segregate across populations may only be partially corrected, while the test statistic of markers that do not segregate may be over-corrected (Marchini et al., 2004; Price et al., 2006). The structured association method splits individuals into different subpopulations, performs association testing within these subgroups, and then merges the evidence for association. A major limitation of this methodology is that only discrete subgroups can be considered. Moreover, its results can vary depending on the selected number of clusters (Price et al., 2006).

Analyses of genotype data in larger cohorts, made possible by the advances in genotyping technology, showed that genome-scale genetic variation could be used to accurately infer population structure (Bauchet et al., 2007; Jakobsson et al., 2008; Li et al., 2008; Tian et al., 2008; Price et al., 2008). In particular, the first principal components (PCs) of the genotype data were shown to correlate with orthogonal geographic axes (Lao et al., 2008; Novembre and Stephens, 2008) and to better distinguish between closely-spaced populations compared to geographic information (Novembre et al., 2008). Price et al. (2006) suggested accounting for population structure by regressing the top (ten) principal components from both genotype and phenotype data prior to performing association testing. An equivalent strategy is to include the leading principle components as covariates within the linear model. The optimal number of PCs can either be selected to minimise inflation (Tian et al., 2008) or be based on the correlation of single PCs with the phenotype (Lee et al., 2011). In analysis of unrelated individuals, twenty principal components are typically sufficient to correct for population structure (Astle and Balding, 2009). However, as the effects of population structure are more severe in analyses of larger cohorts (Marchini et al., 2004), the optimal number of PCs will also depend on the sample size of the considered cohort.

2.3 Genetic analysis with the linear mixed model

A linear mixed model (LMM) describes the outcome variable as a sum of unknown deterministic effects (fixed effects) and unknown random effects. A random effect is the realisation of a random variable of which we model the distribution. Denoting with N, K and Q the number of samples, fixed effects and random effects respectively, a linear mixed model can be cast as

$$\boldsymbol{y} = \underbrace{\boldsymbol{X\beta}}_{\text{fixed effects}} + \underbrace{\boldsymbol{Zb}}_{\text{random effects}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}}, \tag{2.21}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ is the outcome vector, $\boldsymbol{X} \in \mathbb{R}^{N \times K}$ and $\boldsymbol{Z} \in \mathbb{R}^{N \times Q}$ are the design matrix of fixed and random effects respectively, $\boldsymbol{\beta} \in \mathbb{R}^K$ are the fixed effects, $\boldsymbol{b} \sim \mathcal{N}\left(\mathbf{0}, \sigma_b^2 \boldsymbol{\Sigma}\right)$ are the random effects, $\boldsymbol{\Sigma}$ is a known covariance, $\boldsymbol{\psi} \sim \mathcal{N}\left(\mathbf{0}, \sigma_e^2 \boldsymbol{I}_N\right)$ is the residual vector and σ_b^2 and σ_e^2 are the variance parameters of the random effect and noise distributions.

Mixed model equations The joint density of $\boldsymbol{y}$ and $\boldsymbol{b}$ is

$$
\begin{aligned}
p\left(\boldsymbol{y}, \boldsymbol{b} \,|\, \boldsymbol{\beta}, \sigma_e^2, \sigma_b^2\right) &= p\left(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{b}, \sigma_e^2\right) p\left(\boldsymbol{b} \,|\, \sigma_b^2 \boldsymbol{\Sigma}\right) && (2.22)\\
&= \mathcal{N}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}, \sigma_e^2 \boldsymbol{I}_N\right) \mathcal{N}\left(\boldsymbol{b}, \mathbf{0}, \sigma_b^2 \boldsymbol{\Sigma}\right) && (2.23)
\end{aligned}
$$

The values $\boldsymbol{\beta}$ and $\boldsymbol{b}$ that maximise the joint distribution can be obtained by equating the gradients of $f\left(\boldsymbol{\beta}, \boldsymbol{b}\right) = \log p\left(\boldsymbol{y}, \boldsymbol{b} \,|\, \boldsymbol{\beta}, \sigma_e^2, \sigma_b^2\right)$ to 0

$$
\begin{aligned}
\left(\frac{\partial f\left(\boldsymbol{\beta}, \boldsymbol{b}\right)}{\partial \boldsymbol{\beta}}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{b}=\hat{\boldsymbol{b}}} &= 0 && (2.24)\\
\left(\frac{\partial f\left(\boldsymbol{\beta}, \boldsymbol{b}\right)}{\partial \boldsymbol{b}}\right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{b}=\hat{\boldsymbol{b}}} &= 0. && (2.25)
\end{aligned}
$$

Solving the equation system gives the mixed model equations (Henderson, 1950; Henderson et al., 1959)

$$
\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{Z} + \frac{\sigma_e^2}{\sigma_b^2} \boldsymbol{\Sigma}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{b}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{y} \\ \boldsymbol{Z}^\top \boldsymbol{y} \end{bmatrix}. \tag{2.26}
$$

Marginal likelihood As σ_b^2 and σ_e^2 are unknown, one can estimate these variance parameters together with the fixed effects $\boldsymbol{\beta}$ by maximising the marginal likelihood $p\left(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma_e^2, \sigma_b^2\right)$ (Dempster et al., 1981). The marginal likelihood can be obtained by marginalising out the random effect $\boldsymbol{b}$ as follows

$$
\begin{aligned}
p\left(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma_e^2, \sigma_b^2\right) &= \int p\left(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{b}, \sigma_e^2\right) p\left(\boldsymbol{b} \,|\, \sigma_b^2\right) \, \mathrm{d}\boldsymbol{b} && (2.27)\\
&= \mathcal{N}\left(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}, \sigma_b^2 \boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\top + \sigma_e^2 \boldsymbol{I}_N\right). && (2.28)
\end{aligned}
$$

This model is equivalent to Bayesian linear regression. The marginal likelihood is regarded as a function of $\boldsymbol{\beta}$, σ_e^2 and σ_b^2 and it is denoted as $\mathcal{L}\left(\boldsymbol{\beta}, \sigma_g^2, \sigma_e^2\right)$. As MLE of the variance components are biased, parameter inference in LMMs is usually performed maximising the restricted marginal likelihood (see Section A.1).

Contrary to the case of linear regression discussed in Section 2.1.1, there is no close-form solution for the (restricted) MLE of the model parameters. As shown in Section 2.3.3, an efficient derivative-free inference scheme can be used for association testing in univariate models (Lippert et al., 2011). However, the number of variance parameters in variance component analyses and multi-trait models is typically too large for the efficient use of derivative-fee methods. Multiple optimisation schemes have been proposed to optimise the restricted marginal likelihood in these cases, including first-derivative methods, such as expectation maximisation (EM, Dempster et al., 1977) and its improved version PX-EM (Liu et al., 1998; Foulley and Van Dyk, 2000), and second-derivative methods, such as the Newton-Ralphson algorithm (Zhou and Stephens, 2014), the average information REML algorithm (Gilmour et al., 1995) and the Broyden's method (Groeneveld, 1994). In this thesis, we follow (Groeneveld, 1994) and consider the Broyden's method for parameters inference. For a discussion on the different optimisation algorithms for inference in LMMs, I refer to the supplementary information of Loh et al. (2015a).

In the following I give an overview of the different applications of LMMs in genetics.

2.3.1 Accounting for confounding using the linear mixed model

While PC-based approaches can correct for population stratification in human studies of unrelated individuals, such a methodology is less successful to account for more subtle types of confounding. Conversely, linear mixed models have been proven to yield calibrated P values in analyses of model organisms with complex population structure (Yu et al., 2006; Zhao et al., 2007; Kang et al., 2008b) and human GWAS with cryptic related individuals (Kang et al., 2010; Price et al., 2010; Zhou and Stephens, 2012). **Fig. 2.2** shows the QQ plots obtained considering alternative strategies to correct for population structure in a GWAS of flowering time in *Arabidopsis thaliana* (*A. thaliana*). Only the LMM yielded calibrated P values.

The standard LMM used for association testing is

$$\boldsymbol{y} = \underbrace{\boldsymbol{g}\beta}_{\text{variant effect}} + \underbrace{\boldsymbol{X}\boldsymbol{\alpha}}_{\text{covariate effects}} + \underbrace{\boldsymbol{u}}_{\text{confounding}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}} \tag{2.29}$$

where the effects from the genetic variant being tested and the covariates are modelled as fixed effects, while the effect from confounding is modelled as a random vector

$$\boldsymbol{u} \sim \mathcal{N}\left(\mathbf{0}, \sigma_g^2 \boldsymbol{R}\right). \tag{2.30}$$

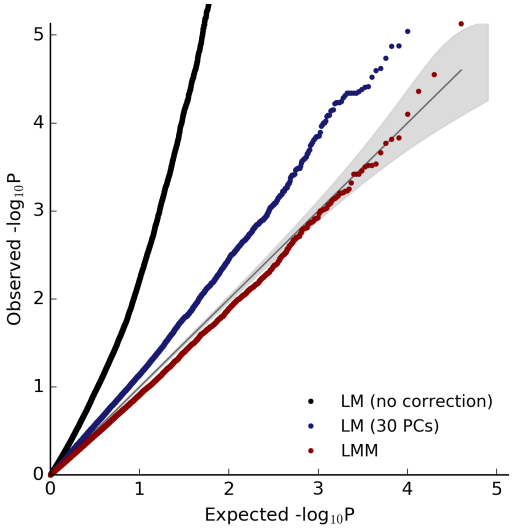


Figure 2.2: **QQ plots obtained from a GWAS of a flowering trait in *A. thaliana* using different correction strategies.** Compared are a linear model (LM) without correction, a linear model with 30 genetic PCs and a linear mixed model (LMM). The shaded area indicates the 99% confidence interval around the diagonal (dark grey line). In particular, I tested for association between of 39,921 SNPs (MAF $> 5\%$) and flowering time at 22 degree Celsious in 192 individuals. Phenotypes have been quantile-normalised to a normal distribution prior to QTL mapping. Data are from Atwell et al. (2010).

Here, $\boldsymbol{R}$ denotes the genetic relatedness matrix, which accounts for the pairwise genetic similarity between individuals. As discussed in the next section, such pair-wise genetic similarities between individuals capture correlations due to both population structure and family relatedness. Marginalising out $\boldsymbol{u}$, the model in (2.29) can be equivalently expressed as

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{g}\beta + \boldsymbol{X}\boldsymbol{\alpha}\,,\, \sigma_g^2\boldsymbol{R} + \sigma_e^2\boldsymbol{I}_N\right). \tag{2.31}$$

This equation suggests an interpretation of the influence of confounding on the phenotype values: confounding induces a "covariance structure" between trait observations. Such a structure is estimated as the genetic relatedness matrix $\boldsymbol{R}$. As we will see in the next section, there are different ways of defining $\boldsymbol{R}$.

2.3.2 Genetic relatedness matrix

Fisher's work (Fisher, 1919) shows that under an additive model with an infinite number of infinitesimal genetic effects, the phenotype is normally distributed and the pheno-

typic correlation between individuals is proportional to the fraction of genetic material that is identical-by-descent (IBD)[5]. Therefore, a natural way of defining the genetic relatedness between two individuals is to use the predicted proportion of the genome that is IBD in the considered pair. Traditionally, these IBD relatedness matrices were estimated from known pedigrees (LANGE et al., 1976). Note that the definition of IBD requires the specification of a base population (i.e. the population of the ancestors, whose average relationship is zero), which in the case of pedigree designs is the population of the founders (Powell et al., 2010).

An alternative that is becoming increasingly common is to estimate relatedness matrices from genome-wide SNPs. SNP-based relatedness matrices have been shown to improve narrow-sense heritability estimates (Visscher et al., 2006; Visscher et al., 2007; Hayes et al., 2009) and to better account for population structure (Kang et al., 2008b; Lee et al., 2010) compared to pedigree-based matrices. Different ways of estimating relatedness matrices from genotype data have been proposed (Oliehoek et al., 2006; Purcell et al., 2007; VanRaden, 2008). A widely-used estimate of the genetic relatedness matrix is the realised relatedness matrix (RRM) (Hayes et al., 2009), which is defined as

$$\boldsymbol{R} = \frac{1}{S}\boldsymbol{G}\boldsymbol{G}^\top, \tag{2.32}$$

where $\boldsymbol{G}$ is the $N \times S$ genotype matrix with standardised genotypes across individuals and S denotes the number of genome-wide variants. Interestingly, the RRM can be obtained from the polygenic model

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{G}\boldsymbol{b}\,,\, \sigma_e^2\boldsymbol{I}_N\right) \tag{2.33}$$

where $\boldsymbol{X}$ is the design matrix of K covariates, $\boldsymbol{\alpha}$ their fixed effects and $\boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}\,,\, \frac{\sigma_g^2}{S}\boldsymbol{I}_S\right)$ are random effects of the genome-wide variants. Note that each genetic marker explains on average variance $\frac{\sigma_g^2}{S}$, so that genome-wide variants jointly explain variance σ_g^2. Marginalising out the random effect we obtain the RRM in one of the terms of the covariance

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha}\,,\, \sigma_g^2 \underbrace{\frac{1}{S}\boldsymbol{G}\boldsymbol{G}^\top}_{\text{RRM}} + \sigma_e^2\boldsymbol{I}_N\right). \tag{2.34}$$

The RRM can also be interpreted as an IBD relatedness matrix where the base population is the current population (Powell et al., 2010).

[5] A locus is IBD in two individuals if it has been inherited by a common ancestor.

Relationship to PC-based methods Principal components of the genotype data $\boldsymbol{G}$ can be calculated as the eigenvectors of the relatedness matrix $\boldsymbol{R} = \frac{1}{S}\boldsymbol{G}\boldsymbol{G}^\top$ (Price et al., 2006). Denoting with $\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$ the eigenvalue decomposition[6] of $\boldsymbol{R}$, the marginalised model in (2.34) can be obtained from the linear mixed model

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha} + \underbrace{\boldsymbol{U}\boldsymbol{b}}_{\text{contribution from PCs}}, \sigma_e^2\boldsymbol{I}_N\right), \quad \text{with } \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_g^2\boldsymbol{S}\right), \tag{2.35}$$

by integrating out the random effects $\boldsymbol{b}$. In summary, while a PC-based linear model only accounts for the first principal components of the genotype data using fixed effects, linear mixed models consider all the principal components using random effects. This explains why LMMs can account for more subtle (i.e. high-rank) confounding compared to PC-based linear models. For further details on the relationship between PC-based models and random effect models I refer to Hoffman (2013).

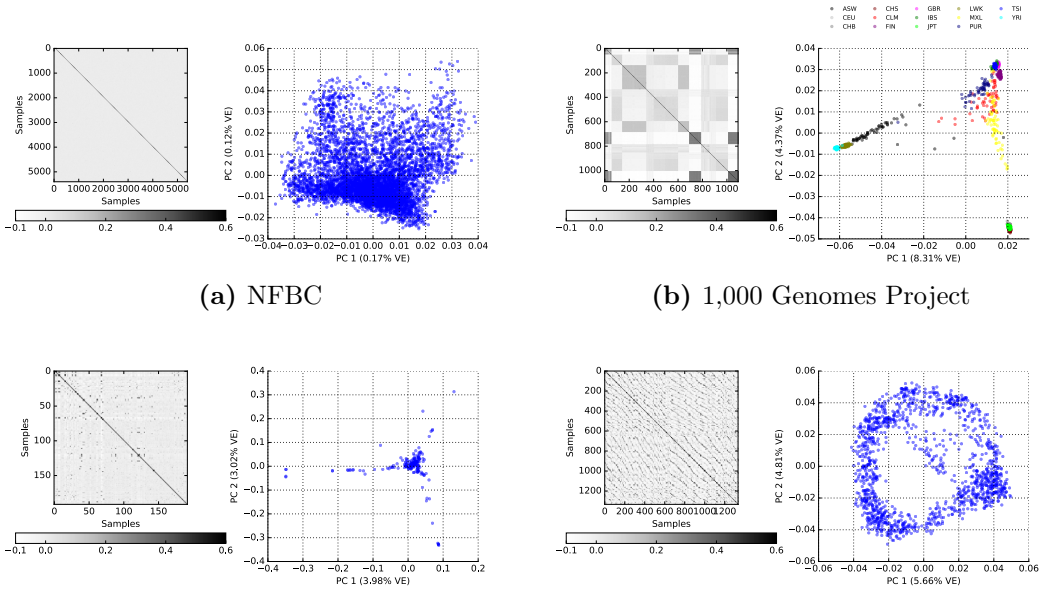
Examples of RRMs **Fig. 2.3** shows RRMs for 4 different cohorts: the Northern Finland Birth Cohort (Sabatti et al., 2009), the Phase 1 1,000 Genomes Project (Goncalo R Abecasis et al., 2012), a cohort of inbred lines of *A. thaliana* (Atwell et al., 2010) and an outbred rat population (Baud et al., 2014). With the exception of the *A. thaliana* dataset, which I considered to generate **Fig. 2.2**, these datasets will be considered in the next chapters.

Modelling relationships with MAF, LD tagging and other properties. The polygenic model considered above builds on the prior assumption that every variant equally contributes to the phenotypic variance. However, the same framework can be employed to model relationships of these contributions with variant-specific properties, such as minor allele frequency, LD tagging and genotype imputation quality scores (Speed et al., 2016). This can be derived from the generative linear model in Eq (2.33) by considering a variant-specific prior on effect sizes, $\beta_s \sim \mathcal{N}\left(0, \sigma_s^2\right)$, where the variance σ_s^2 is a function of such properties. For example, following Speed et al. (2016), a relationship with the minor allele frequency can be introduced by setting

$$\sigma_s^2 = \frac{[f_s(1-f_s)]^\alpha}{\sum_k [f_k(1-f_k)]^\alpha}, \tag{2.36}$$

[6]The columns of $\boldsymbol{U}$ are the eigenvectors while the diagonal entries of $\boldsymbol{S}$ are the corresponding eigenvalues. Eigenvectors and eigenvalues are ordered in decreasing order of variance explained.

where f_s is the minor allele frequency of variant s and α defines the relationship between the expected variance and the minor allele frequency. Specifically, $\alpha < 0$ corresponds to the assumption that low-frequency variants have stronger contributions in comparison to common variants while $\alpha > 0$ corresponds to the converse assumption. The uniform model can be recovered setting $\alpha = 0$. Note that different assumptions on per-variant contributions correspond to different strategies of rescaling the columns of $\boldsymbol{G}$ prior to building the RRM (see Eq (2.32)). Unless stated otherwise, I will make the assumption of uniform contributions, which corresponds to taking $\boldsymbol{G}$ with standardised columns in Eq (2.32).



(a) NFBC **(b)** 1,000 Genomes Project

(c) *A. thaliana* dataset **(d)** Rat dataset

Figure 2.3: **Realised relatedness matrices for four different cohorts.** Shown are the realised relatedness matrices and the scatter plots of the first two principal components for four different cohorts. (**a**) 5,402 unrelated individuals from the Northern Finland Birth Cohort (NFBC, phs000276.v1.p1) (Sabatti et al., 2009). (**b**) 1,092 individuals from the 14 populations that are part of the Phase 1 1,000 Genomes Project (Goncalo R Abecasis et al., 2012). (**c**) 192 inbred Arabidopsis lines from Atwell et al. (2010) used in **Fig. 2.2**. (**d**) 1,334 rats from the outbred population in Baud et al. (2014).

2.3.3 Efficient linear mixed models for GWAS

Computations associated with inference in linear mixed models scale cubically with the number of individuals in the dataset. For example, the evaluation of the restricted

marginal likelihood in Eq (A.3) entails the computation of operations with $O(N^3)$ complexity, such the matrix inverse and of the log determinant of the total covariance. However, for the model in (2.29) it is possible to speed up computations (Kang et al., 2008b; Kang et al., 2010; Lippert et al., 2011; Zhou et al., 2012) enabling applications to large cohorts. These strategies reduce the computational complexity from $O(N^3)$ per-variant to a single $O(N^3)$ cost up-front and a per-variant complexity of $O(N^2)$. The complexity can be further reduced to $O(N^2)$ for the up-front computation and a per-test complexity of $O(N)$, provided the genetic relatedness matrix is low-rank. In practice, this can be achieved through a feature selection approach, selecting a small proportion of all genome-wide variants to estimate $\boldsymbol{R}$ (Listgarten et al., 2012) or by randomly selecting a subset of the genome-wide variants. In the following I will briefly describe the efficient FaST-LMM algorithm proposed by Lippert et al. (2011).

The objective is to optimise the log marginal likelihood

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{g}\beta + \boldsymbol{X}\boldsymbol{\alpha}\,,\, \sigma_g^2\boldsymbol{R} + \sigma_e^2\boldsymbol{I}_N\right) \tag{2.37}$$

iteratively for different tested variants (i.e.,for different genotypes vectors $\boldsymbol{g}$). The strategy can be summarised in three steps, (1) eigenvalue decomposition of the genetic similarity matrix, (2) transformation of the data in a space where they are uncorrelated and (3) single-variant testing in the new space.

1. Eigenvalue decomposition of $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$ needs to be computed only once upfront and has complexity $O(N^3)$.

2. Introducing $\delta = \sigma_e^2/\sigma_g^2$ and using $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$ we can write

$$\boldsymbol{K} = \sigma_g^2\left(\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top + \delta\boldsymbol{I}_N\right) = \sigma_g^2\boldsymbol{U}\left(\boldsymbol{S} + \delta\boldsymbol{I}_N\right)\boldsymbol{U}^\top. \tag{2.38}$$

The log marginal likelihood of the model is

$$\begin{aligned}\log\mathcal{L}\left(\beta, \boldsymbol{\alpha}, \sigma_g^2, \delta\right) &= -\frac{N}{2}\log\left(2\pi\right) - \frac{1}{2}\log\det\boldsymbol{K} + & (2.39)\\ &\quad -\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{g}\beta - \boldsymbol{X}\boldsymbol{\alpha}\right)^\top\boldsymbol{K}^{-1}\left(\boldsymbol{y} - \boldsymbol{g}\beta - \boldsymbol{X}\boldsymbol{\alpha}\right) & (2.40)\end{aligned}$$

Using that

$$\boldsymbol{K}^{-1} = \frac{1}{\sigma_g^2}\boldsymbol{U}\underbrace{(\boldsymbol{S}+\delta\boldsymbol{I}_N)^{-1}}_{\boldsymbol{D}_\delta}\boldsymbol{U}^\top \tag{2.41}$$

$$\log\det\boldsymbol{K} = N\log\sigma_g^2 + \log\det(\boldsymbol{S}+\delta\boldsymbol{I}_N) \tag{2.42}$$

where we introduced $\boldsymbol{D}_\delta = (\boldsymbol{S}+\delta\boldsymbol{I}_N)^{-1}$, we have

$$\log\mathcal{L}\left(\beta,\boldsymbol{\alpha},\sigma_g^2,\delta\right) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log\sigma_g^2 + \frac{1}{2}\log\det\boldsymbol{D}_\delta +$$

$$-\frac{1}{2\sigma_g^2}\left(\boldsymbol{U}^\top\boldsymbol{y} - \boldsymbol{U}^\top\boldsymbol{g}\beta - \boldsymbol{U}^\top\boldsymbol{X}\boldsymbol{\alpha}\right)^\top\boldsymbol{D}_\delta\left(\underbrace{\boldsymbol{U}^\top\boldsymbol{y}}_{\tilde{\boldsymbol{y}}} - \underbrace{\boldsymbol{U}^\top\boldsymbol{g}}_{\tilde{\boldsymbol{g}}}\beta - \underbrace{\boldsymbol{U}^\top\boldsymbol{X}}_{\tilde{\boldsymbol{X}}}\boldsymbol{\alpha}\right) =$$

$$-\frac{N}{2}\log(2\pi) - \frac{1}{2}\log\sigma_g^2 + \frac{1}{2}\log\det\boldsymbol{D}_\delta +$$

$$-\frac{1}{2\sigma_g^2}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{g}}\beta - \tilde{\boldsymbol{X}}\boldsymbol{\alpha}\right)^\top\boldsymbol{D}_\delta\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{g}}\beta - \tilde{\boldsymbol{X}}\boldsymbol{\alpha}\right). \tag{2.43}$$

Here, $\tilde{\boldsymbol{y}}$, $\tilde{\boldsymbol{g}}$ and $\tilde{\boldsymbol{X}}$ denote the "rotated" phenotype vector, genotype matrix and covariate matrix, respectively. This rotation needs to be performed for each genotype $\boldsymbol{g}$ to be tested, which is equivalent to rotating the full genotype matrix. The rotation of the phenotype vector, the genotype data and the matrix of covariates requires $O(N^2 + N^2S + N^2K)$ operations, where K and S denote the numbers of covariates and variants respectively.

3. Fixing delta and regarding the likelihood in Eq (2.44) as a function of only β, $\boldsymbol{\alpha}$ and σ_g^2 we have

$$\log\mathcal{L} = \text{const} - \frac{1}{2}\log\sigma_g^2 - \frac{1}{2\sigma_g^2}\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{g}}\beta - \tilde{\boldsymbol{X}}\boldsymbol{\alpha}\right)^\top\boldsymbol{D}_\delta\left(\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{g}}\beta - \tilde{\boldsymbol{X}}\boldsymbol{\alpha}\right) =$$

$$\text{const} - \frac{1}{2}\log\sigma_g^2 - \frac{1}{2\sigma_g^2}\left(\boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{y}} - \boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{g}}\beta - \boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{X}}\boldsymbol{\alpha}\right)^\top\left(\boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{y}} - \boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{g}}\beta - \boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{X}}\boldsymbol{\alpha}\right)$$

and the MLE of β, $\boldsymbol{\alpha}$ and σ_g^2 are the MLE of the linear model[7]

$$\boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{y}} \sim \mathcal{N}\left(\boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{g}}\beta + \boldsymbol{D}_\delta^{\frac{1}{2}}\tilde{\boldsymbol{X}}\boldsymbol{\alpha}\,,\,\sigma_g^2\boldsymbol{I}_N\right), \tag{2.44}$$

which can be computed in closed-form. Optimisation can thus proceed by con-

[7] the two likelihood functions differ by a constant.

sidering a 1D Brent search optimisation routine in δ (Brent, 1971) while using close-form computations for β, $\boldsymbol{\alpha}$ and σ_g^2. Note that all operations at this step are linear in N.

The same strategy can be used to calculate the REML of the model (see Eq (A.3))

$$\log \mathcal{L}_{\text{REML}}\left(\beta, \boldsymbol{\alpha}, \sigma_g^2, \delta\right) = -\frac{N-K}{2}\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{K} - \log\det \boldsymbol{A} \quad (2.45)$$

$$-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{g}\beta - \boldsymbol{X}\boldsymbol{\alpha}\right)^\top \boldsymbol{K}^{-1}\left(\boldsymbol{y} - \boldsymbol{g}\beta - \boldsymbol{X}\boldsymbol{\alpha}\right) \quad (2.46)$$

where $\boldsymbol{A} = \boldsymbol{X}^\top \boldsymbol{K}^{-1}\boldsymbol{X}$. Note that as we are interested in testing for β the genotype vector $\boldsymbol{g}$ is not considered in the regularisation term $\log\det \boldsymbol{A}$ (Lippert et al., 2011; Zhou et al., 2012).

2.3.4 Variance component models

We have seen how genetic relatedness matrices can be used in linear mixed models to account for confounding effects in GWAS. The central idea of this approach is to estimate the structure of confounding from genetic relatedness between individuals. However, in analyses of datasets where environmental factors are controlled for, the same models can be used to estimate narrow-sense heritability. Traditionally, LMMs have been considered for heritability estimation in genetic analysis of animal models using either pedigree-based or SNP-based relatedness matrices (Lynch and Walsh, 1998; Valdar et al., 2006; The Rat Genome Sequencing and Mapping Consortium, 2013). More recently, this approach has been applied to human cohorts of unrelated samples (Yang et al., 2010; Lee et al., 2012b; Gusev et al., 2014), where population structure can be accounted for by using the top genetic principal components (Yang et al., 2011a). In this section, I briefly review the concepts and models for heritability estimation and variance decomposition. The standard software tool for these analyses is GCTA (Genome-wide Complex Trait Analysis, Yang et al. (2011a)). GCTA employs gradient-based parameter inference using the PX-AI algorithm (Meyer, 2006), which combines EM and average information REML.

Estimating narrow-sense heritability In absence of confounding, if genotype data at causal variants were known, one could estimate narrow sense heritability using the LMM

$$\boldsymbol{y} = \mathbf{1}_N\mu + \boldsymbol{G}^{(\mathrm{c})}\boldsymbol{u}^{(\mathrm{c})} + \boldsymbol{\psi}, \quad \boldsymbol{u}^{(\mathrm{c})} \sim \mathcal{N}\left(\mathbf{0},\, \frac{\sigma_g^2}{S_c}\boldsymbol{I}_{S_c}\right), \quad \boldsymbol{\psi} \sim \mathcal{N}\left(\mathbf{0},\, \sigma_e^2\boldsymbol{I}_N\right), \quad (2.47)$$

where $\mathbf{1}_N\mu$ is a mean term, S_c is the number of causal variants, $\boldsymbol{G}^{(\mathrm{c})}$ is the standardised genotype matrix at causal variants and σ_g^2 is the variance explained jointly by all causal variants[8]. Marginalising out the random effects we have

$$\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{1}\mu, \sigma_g^2 \boldsymbol{R}^{(\mathrm{c})} + \sigma_e^2 \boldsymbol{I}_N\right), \tag{2.48}$$

where $\boldsymbol{R}^{(\mathrm{c})} = \frac{1}{S_c}\boldsymbol{G}^{(\mathrm{c})}\boldsymbol{G}^{(\mathrm{c})\top}$ is the RRM at causal variants. Using this model, the narrow-sense heritability of the trait can be estimated as

$$h^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2}, \tag{2.49}$$

where $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ are the restricted MLE of σ_g^2 and σ_e^2 respectively. If covariates $\boldsymbol{X}$ are included in the model, we have

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha}, \sigma_g^2 \boldsymbol{R}^{(\mathrm{c})} + \hat{\sigma}_e^2 \boldsymbol{I}_N\right) \tag{2.50}$$

and

$$h^2 = \frac{\hat{\sigma}_g^2}{\mathrm{var}\left(\boldsymbol{X}\hat{\boldsymbol{\alpha}}\right) + \hat{\sigma}_g^2 + \hat{\sigma}_e^2} \tag{2.51}$$

where $\hat{\boldsymbol{\alpha}}$, $\hat{\sigma}_g^2$ and $\hat{\sigma}_e^2$ are the MLE of $\boldsymbol{\alpha}$, σ_g^2 and σ_e^2 respectively.

As causal variants are generally not known, one can use the model with an RRM estimated from all genotyped variants ($\boldsymbol{R}$) to estimate narrow-sense heritability

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha}, \sigma_g^2 \boldsymbol{R} + \sigma_e^2 \boldsymbol{I}_N\right) \tag{2.52}$$

However, as a consequence of the incomplete and uneven tagging of causal variants from the observed SNPs, this model will give lower heritability estimates compared to the ideal model. Yang et al. (2010) studied the effect of the incomplete tagging of causal variants when observed and causal variants have different MAF distributions. To alleviate this bias, they proposed a shrunk version of the RRM obtained by regressing $\boldsymbol{R}$ towards the identity matrix (Powell et al., 2010). Speed et al. (2012) studied the effect of uneven tagging of causal variants caused by the different extent of LD along the genome and found that it can lead to biased heritability estimates. Therefore they proposed an adjusted RRM where SNPs are weighted according to local LD.

[8]Each variant explains variant $\frac{\sigma_g^2}{S_c}$ on average.

Variance decomposition The approach presented above can be extended to partition the phenotypic variance across different sets of SNPs, for example SNPs from different genetic regions (Yang et al., 2011b; Lee et al., 2013) or SNPs with distinct functional categories (Gusev et al., 2014). Formally, let $\{\boldsymbol{G}_1, \ldots, \boldsymbol{G}_M\}$ denote the standardised genotype values corresponding to M disjunct sets of variants. Additionally, let S_m denote the number of variants in set m and $\boldsymbol{R}_m = \frac{1}{S_m}\boldsymbol{G}_m\boldsymbol{G}_m^\top$ denote the RRM estimated from the variants in set m. The variance explained by the different sets can be obtained from the model

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha},\ \sum_{m=1}^{M}\sigma_m^2\boldsymbol{R}_m + \sigma_e^2\boldsymbol{I}_N\right) \tag{2.53}$$

as

$$h_m^2 = \frac{\hat{\sigma}_m^2}{\operatorname{var}(\boldsymbol{X}\hat{\boldsymbol{\alpha}}) + \sum_{m=1}^{M}\hat{\sigma}_m^2 + \hat{\sigma}_e^2} \tag{2.54}$$

where $\hat{\boldsymbol{\alpha}}$, $\hat{\sigma}_m^2$ and $\hat{\sigma}_e^2$ are the MLE of $\boldsymbol{\alpha}$, σ_m^2 and σ_e^2 respectively.

Alternatively, in order to estimate the variance explained by a specific set of genetic variants, Kostem and Eskin (2013) proposed the two-variance-component model

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\alpha},\ \sigma_p^2\boldsymbol{R}_p + \sigma_g^2\boldsymbol{R}_g + \sigma_e^2\boldsymbol{I}_N\right), \tag{2.55}$$

where $\boldsymbol{R}_p$ is the partial RRM based on the variants in the set while $\boldsymbol{R}_g$ is the global RRM built from genome-wide variants. This model estimates the variance explained by the SNP-set while accounting for polygenic effects from genome-wide variants and, in case of confounding, for population structure and cryptic relatedness.

Standard errors on variance component estimates can be calculated using the Fisher information matrix as discussed in Section D.3.

2.3.5 Set tests

Set tests are a class of statistical models that enable testing for association between a set of genetic variants and the trait of interest. Set tests can help reduce the multiple hypothesis testing burden and detect loci harbouring multiple causal alleles (Listgarten et al., 2013).

Depending on the strategy employed to aggregate effects from multiple variants, set tests can be broadly classified in three categories (Schifano et al., 2012): methods that combine results from single-variant association testing (Conneely and Boehnke,

2007; Gao et al., 2008; Liu et al., 2010; Moskvina and Schmidt, 2008), methods that consider association testing with a weighted sum of the set genotypes (Li et al., 2009; Gauderman et al., 2007) and methods that use genetic similarity matrices (Tzeng et al., 2003; Schaid et al., 2005; Wessel and Schork, 2006; Tzeng et al., 2009; Mukhopadhyay et al., 2010; Wu et al., 2010; Schifano et al., 2012; Listgarten et al., 2013; Chen et al., 2013; Lippert et al., 2014a). The methods in the last category can be formulated within the LMM framework, where the effects of the variants in the set are modelled as random.

The first set tests have considered linear mixed models with an only variance component to jointly model the effects of the variants in the set (Wu et al., 2011; Wu et al., 2010). More recent implementations have introduced a second variance component in the model to account for relatedness (Listgarten et al., 2013; Lippert et al., 2014a; Schifano et al., 2012; Chen et al., 2013). These set tests with two variance components consider the model in Eq (2.55) and test for $\sigma_p^2 \neq 0$. As discussed in Section 3.1.2, for variance component tests the assumptions underlying Wilk's theorem do not hold and as a consequence, the asymptotic distribution of the LLR test statistic is generally not known (Molenberghs and Verbeke, 2007; Self and Liang, 1987; Dominicus et al., 2006). For this reason, a large class of set tests employ score-based test statistics, whose asymptotic distribution is known. Score-based set tests have been used for testing for associations with rare and common variants (Wu et al., 2011; Wu et al., 2010). As an alternative to score-based set tests, Listgarten et al. (2013) have proposed an efficient procedure based on permutations to compute P values from the likelihood ratio test statistic. LLR-based tests have been shown to be more powered than score-based tests in analyses of real data (Lippert et al., 2014a). Importantly, as we will see in detail in the next chapters, the computational efficiency of both score-based and LLR-based set tests strongly depends on the number of variants in the considered genomic region. In particular, for regions for which the number of variants is greater than the number of individuals, computations scale cubically with the number of individuals.

2.3.6 Genomic predictions

Suppose we have observed phenotype data and K covariates for N individuals (in-sample individuals) and the genetic relatedness matrix for a larger set of $N + N_\star$ individuals. The phenotypes of the non-phenotyped $N^\star$ individuals (out-of-sample individuals) can be predicted using the genetic relatedness matrix. Let $\boldsymbol{y} \in \mathbb{R}^N$ denote the phenotype vector of the in-sample individuals and $\boldsymbol{R}^{(\text{all})} \in \mathbb{R}^{(N+N_\star)\times(N+N_\star)}$ denote the genetic relatedness matrix of all $N + N_\star$ individuals. We can order individuals in

$\boldsymbol{R}^{(\text{all})}$ such that

$$\boldsymbol{R}^{(\text{all})} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{R}_\star^\top \\ \boldsymbol{R}_\star & \boldsymbol{R}_{\star\star} \end{bmatrix}, \tag{2.56}$$

where $\boldsymbol{R} \in \mathbb{R}^{N\times N}$ is the relatedness matrix for in-sample individuals, $\boldsymbol{R}_{\star\star} = \frac{1}{S}\boldsymbol{G}_\star\boldsymbol{G}_\star^\top \in \mathbb{R}^{N_\star\times N_\star}$ is the relatedness matrix for out-of-sample individuals and $\boldsymbol{R}_\star \in \mathbb{R}^{N_\star\times N}$ is the cross covariance between out-of-sample and in-sample individuals. Denoting with $\boldsymbol{u}_\star$ the genomic prediction vector for the out-of-sample individuals, we can model the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{u}_\star$ as

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u}_\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{X}\boldsymbol{\alpha} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_g^2\boldsymbol{R} + \sigma_e^2\boldsymbol{I_N} & \sigma_g^2\boldsymbol{R}_\star^\top \\ \sigma_g^2\boldsymbol{R}_\star^\top & \sigma_g^2\boldsymbol{R}_{\star\star} \end{bmatrix} \right), \tag{2.57}$$

where $\boldsymbol{X} \in \mathbb{R}^{N\times K}$ is the design matrix of the K covariates for in-sample individuals. Using the rule for conditioning Gaussians, we have

$$\boldsymbol{u}_\star \,|\, \boldsymbol{y} \sim \mathcal{N}\left(\, \boldsymbol{\mu}_u \,,\, \boldsymbol{\Sigma}_u \,\right) \tag{2.58}$$

where

$$\boldsymbol{\mu}_u = \underbrace{\sigma_g^2\boldsymbol{R}_\star^\top \left(\sigma_g^2\boldsymbol{R} + \sigma_e^2\boldsymbol{I_N}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha})}_{\text{BLUP}} \tag{2.59}$$

$$\boldsymbol{\Sigma}_u = \sigma_g^2\boldsymbol{R}_{\star\star} - \sigma_g^2\boldsymbol{R}_\star^\top \left(\sigma_g^2\boldsymbol{R} + \sigma_e^2\boldsymbol{I_N}\right)^{-1}\boldsymbol{R}_\star. \tag{2.60}$$

The genetic parameters can be estimated by maximising the marginal likelihood of $\boldsymbol{y}$. Subsequently, Eq (2.59) can be used for predicting phenotypes for out-of-sample individuals. The predictor in Eq (2.59) is known as Best Linear Unbiased Predictor (BLUP) and has been extensively used in livestock breeding (Henderson, 1984; Lee et al., 2008; Clark and Werf, 2013).

If covariates are observed also for out-of-sample individuals, they can also be used for predicting out-of-sample. Denoting with $\boldsymbol{X}_\star \in \mathbb{R}^{N_\star\times K}$ the design matrix of the covariates for out-of-sample individuals, the predictions from genetic relatedness and covariates follows the conditional distribution

$$\boldsymbol{u}_\star^{(\text{cov})} \,|\, \boldsymbol{y} \sim \mathcal{N}\left(\, \boldsymbol{\mu}_u^{(\text{cov})} \,,\, \boldsymbol{\Sigma}_u \,\right), \tag{2.61}$$

where

$$\boldsymbol{\mu}_u^{(\text{cov})} = \boldsymbol{X}_\star\boldsymbol{\alpha} + \sigma_g^2\boldsymbol{R}_\star^\top \left(\sigma_g^2\boldsymbol{R} + \sigma_e^2\boldsymbol{I_N}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\alpha}). \tag{2.62}$$

2.4 Extension to the analysis of multiple traits

Models for the joint analysis of multiple traits can be broadly classified in four groups: methods based on principal components (Aschard et al., 2014), canonical correlation analysis (Ferreira and Purcell, 2009), multivariate linear regression (Bottolo et al., 2013), and meta analysis (Bolormaa et al., 2014). Multi-trait LMMs are a class of multivariate linear regression models and have been used to estimate genetic and environmental correlations as well as for association testing, first in the field of animal breeding (Henderson, 1984; Jiang and Zeng, 1995), and more recently in human genetics (Lee et al., 2012a; Korte et al., 2012). In association testing, the multi-trait LMM has been shown to improve statistical power by leveraging genetic and environmental correlations between traits (Korte et al., 2012; Zhou and Stephens, 2014). Additionally, these class of models can also enhance the interpretation of genetic associations by testing for effects that are either shared across all of the considered traits or specific to some (Korte et al., 2012). Recently, computational advances have enabled application of these models to cohorts of thousands of individuals and multiple traits (Rakitsch et al., 2013; Zhou and Stephens, 2014; Lippert et al., 2014c; Furlotte and Eskin, 2015).

In Section 2.4.1, I define some basic operators and distributions that are central to multivariate modelling. In Section 2.4.2, I introduce the matrix-variate mixed model that is used for multi-trait analysis. In Section 2.4.3, I discuss joint statistical testing across multiple traits while, in Section 2.4.4, I give an overview of the algebraic speed-ups that enable applications to larger cohorts.

2.4.1 Mathematical background

In this section, I introduce some basic notation and concepts for multivariate modelling.

Definition. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two matrices having dimensions $M \times N$ and $Q \times R$ respectively, the Kronecker product of the two matrices $\boldsymbol{A} \otimes \boldsymbol{B}$ has dimensions $MQ \times NR$ and is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} A_{11}\boldsymbol{B} & \dots & A_{1N}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ A_{M1}\boldsymbol{B} & \dots & A_{MN}\boldsymbol{B} \end{pmatrix}. \tag{2.63}$$

Definition. Let $\boldsymbol{A}$ be an $M \times N$ matrix, $\mathrm{vec}\,(A)$ concatenates its columns into an MN-dimensional vector.

Properties

- $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$ (2.64)
- $(\boldsymbol{A} \otimes \boldsymbol{B}) \operatorname{vec}(\boldsymbol{X}) = \operatorname{vec}\left(\boldsymbol{BXA}^{\top}\right)$ (2.65)

Definition. A $N \times P$ matrix $\boldsymbol{X}$ follows a matrix-variate normal distribution

$$\boldsymbol{X} \sim \mathrm{MVN}(\boldsymbol{M},\ \boldsymbol{C},\ \boldsymbol{R}) \tag{2.66}$$

if and only if

$$\operatorname{vec}(\boldsymbol{X}) \sim \mathcal{N}(\operatorname{vec}(\boldsymbol{M}),\ \boldsymbol{C} \otimes \boldsymbol{R}). \tag{2.67}$$

2.4.2 The matrix-variate linear mixed model

Let us consider a population dataset consisting of N individuals and P traits. Further, let $\boldsymbol{y}_p \in \mathbb{R}^N$ denote the phenotype vector of trait p and $\boldsymbol{X} \in \mathbb{R}^{N \times K}$ the design matrix for K covariates. For each trait we can assume the polygenic model

$$\boldsymbol{y}_p \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{b}_p, \sigma^2_{g_p}\boldsymbol{R} + \sigma^2_{e_p}\boldsymbol{I}_N\right), \tag{2.68}$$

where $\boldsymbol{b}_p \in \mathbb{R}^K$ denotes the effects of the covariates, $\boldsymbol{\psi}_p \in \mathbb{R}^N$ is the residual vector for trait p, $\boldsymbol{R}$ is the $N \times N$ genetic relatedness matrix and $\sigma^2_{g_p}$ and $\sigma^2_{e_p}$ are the genetic and residual variance components for trait p. Multi-trait LMMs model the covariance between traits p and p' as (Henderson, 1984)

$$\operatorname{Cov}\left(\boldsymbol{y}_p, \boldsymbol{y}'_p\right) = \rho_{g_{pp'}}\sigma_{g_p}\sigma_{g_{p'}}\boldsymbol{R} + \rho_{n_{pp'}}\sigma_{e_p}\sigma_{e_{p'}}\boldsymbol{I}_N. \tag{2.69}$$

where $\rho_{g_{pp'}}$ and $\rho_{n_{pp'}}$ denote the genetic and environmental correlations between traits p and p', respectively.

Introducing

$$\boldsymbol{Y} = \begin{bmatrix}\boldsymbol{y}_1 & \cdots & \boldsymbol{y}_P\end{bmatrix} \in \mathbb{R}^{N \times P}, \tag{2.70}$$

$$\boldsymbol{B} = \begin{bmatrix}\boldsymbol{b}_1 & \cdots & \boldsymbol{b}_P\end{bmatrix} \in \mathbb{R}^{N \times K}, \tag{2.71}$$

$$\boldsymbol{\Psi} = \begin{bmatrix}\boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_P\end{bmatrix} \in \mathbb{R}^{N \times P}, \tag{2.72}$$

and defining the genetic and environmental covariances as

$$\boldsymbol{C}_g = \begin{bmatrix} \sigma^2_{g1} & \rho_g\sigma_{g1}\sigma_{g2} & \cdots & \rho_g\sigma_{g1}\sigma_{gP} \\ \rho_g\sigma_{g1}\sigma_{g2} & \sigma^2_{g2} & \cdots & \rho_g\sigma_{g2}\sigma_{gP} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_g\sigma_{g1}\sigma_{gP} & \rho_g\sigma_{g2}\sigma_{gP} & \cdots & \sigma^2_{gP} \end{bmatrix}, \tag{2.73}$$

$$\boldsymbol{C}_n = \begin{bmatrix} \sigma^2_{n1} & \rho_n\sigma_{n1}\sigma_{n2} & \cdots & \rho_n\sigma_{n1}\sigma_{nP} \\ \rho_n\sigma_{n1}\sigma_{n2} & \sigma^2_{n2} & \cdots & \rho_n\sigma_{n2}\sigma_{np} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_n\sigma_{n1}\sigma_{nP} & \rho_n\sigma_{n2}\sigma_{nP} & \cdots & \sigma^2_{nP} \end{bmatrix}, \tag{2.74}$$

the model can be specified as the matrix-variate mixed model

$$\boldsymbol{Y} = \boldsymbol{F}\boldsymbol{B} + \boldsymbol{U} + \boldsymbol{\Psi}, \tag{2.75}$$

where $\boldsymbol{U}$ and $\boldsymbol{\Psi}$ follow matrix-variate normal distributions

$$\boldsymbol{U} \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{R}, \boldsymbol{C}_g\right) \tag{2.76}$$

$$\boldsymbol{\Psi} \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{I}_N, \boldsymbol{C}_n\right). \tag{2.77}$$

Using the definition of matrix-variate normal and the properties of the Kronecker product and the vec operator, the marginal likelihood of the model can be equivalently expressed as

$$\mathcal{L}\left(\boldsymbol{B}, \boldsymbol{C}_g, \boldsymbol{C}_n\right) = \mathcal{N}\left(\text{vec}\left(\boldsymbol{Y}\right) | \text{vec}\left(\boldsymbol{F}\boldsymbol{B}\right), \boldsymbol{C}_g \otimes \boldsymbol{R} + \boldsymbol{C}_n \otimes \boldsymbol{I}_N\right), \tag{2.78}$$

which we want to maximise with respect to $\boldsymbol{B}$, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$. As a general $P \times P$ covariance matrix is described by $\frac{1}{2}P(P+1)$ parameters (see Section D.2), the model in (2.91) has $P(P+1+K)$ parameters. Given the large number of parameters, derivative-free methods (as the approach discussed in Section 2.3.3) are highly inefficient. Consequently, the likelihood of multi-trait LMMs is commonly optimised using a gradient-based optimisation algorithm. For example, Korte et al. (2012) considered average information REML (Gilmour et al., 1995) while Zhou and Stephens (2014) proposed Newton-Raphson (Thompson, 1973) in combination with the PX-EM (Liu et al., 1998; Foulley and Van Dyk, 2000). An exception is the work from Furlotte and Eskin (2015), where the authors proposed to first estimate the marginal variances using a single-trait model and subsequently, to estimate correlations of pairs of traits using a grid-search

approach.

2.4.3 Association testing

Denoting with $\boldsymbol{g} \in \mathbb{R}^N$ the genotype of the variant being tested and $\boldsymbol{b} \in \mathbb{R}^P$ its effects on the P traits, the standard linear mixed model for association testing across multiple traits is

$$\boldsymbol{Y} = \underbrace{\boldsymbol{g}\boldsymbol{b}^\top}_{\text{genetic variant}} + \underbrace{\boldsymbol{F}\boldsymbol{B}}_{\text{covariates}} + \underbrace{\boldsymbol{U}}_{\text{confounding}} + \underbrace{\boldsymbol{\Psi}}_{\text{noise}} \tag{2.79}$$

where $\boldsymbol{F} \in \mathbb{R}^{N \times K}$ and $\boldsymbol{B} \in \mathbb{R}^{K \times P}$ the design matrix of K covariates and their effect sizes respectively, and $\boldsymbol{U}$ and $\boldsymbol{\Psi}$ are respectively the effect from confounding and the residual and follow matrix-variate normal distributions:

$$\boldsymbol{U} \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{R}, \boldsymbol{C}_g\right), \quad \boldsymbol{\Psi} \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{I}_N, \boldsymbol{C}_n\right) \tag{2.80}$$

In (4.2) $\boldsymbol{R} \in \mathbf{R}^{N \times N}$ denotes the RRM, $\boldsymbol{C}_g$ is a $P \times P$ covariance matrix describing covariances between traits due to confounding/polygenic effects while $\boldsymbol{C}_n$ is the residual $P \times P$ trait covariance. While $\boldsymbol{R}$ is estimated from genotype data, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ need to be estimated by (restricted) maximum likelihood.

Korte et al. (2012) proposed using the model in (2.79) to test for different hypothesis on the design of the genetic effect across traits. Introducing a common effect model, where the variant has the same effect size across all traits ($\boldsymbol{b} = \boldsymbol{1}_P b$), and a no-effect model, where $\boldsymbol{b} = \boldsymbol{0}_P$, we can define the following tests:

- *Any effect test*: P degree-of-freedom (dof) test where the model in (2.79) is tested against the no-association model (i.e. $\boldsymbol{b} \neq \boldsymbol{0}_P$);

- *Common effect test*: 1 dof test where the common effect model is tested against the no-association model (i.e. $\boldsymbol{b} = \boldsymbol{1}b$ and $b \neq 0$);

- *Specific effect test*: this is a 1 dof test where a model that contains both a common effect across all traits and a specific effect for trait p is tested against the common effect model (i.e. $\boldsymbol{b} = \boldsymbol{1}_P b_c + \mathbb{I}_p b_s$ with $(\mathbb{I}_p)_i = \delta_{ip}$, and $b_s \neq 0$).

Similarly to univariate testing P values can be obtained using a LLR test. In Section 5.2.2, we will generalise this testing framework to consider more complex designs.

2.4.4 Efficient implementation

Fitting the no-association model

Let us start considering the no-association model, Eq (2.78) with $\beta = 0$. As discussed above, model parameters are typically optimised using a gradient-based optimisation algorithm, which requires multiple evaluations of the (restricted) marginal likelihood and its gradients. A naive approach to compute these quantities scales with the cube of the number of individuals and traits, hindering application to bigger cohorts. However, efficient algebraic implementations can reduce computational complexity to an up-front $O(N^3)$ operation, after which gradient-based optimisation proceeds linearly in N (Rakitsch et al., 2013; Zhou and Stephens, 2014; Lippert et al., 2014c; Furlotte and Eskin, 2015). These speedups are possible because of the particular structure of the total covariance matrix $\boldsymbol{A}$, which is the sum of two Kronecker products

$$\boldsymbol{A} = \boldsymbol{C}_g \otimes \boldsymbol{R}_g + \boldsymbol{C}_n \otimes \boldsymbol{I}_N. \tag{2.81}$$

Following Rakitsch et al. (2013) and using the notation $\boldsymbol{M} = \boldsymbol{U}_{\boldsymbol{M}} \boldsymbol{S}_{\boldsymbol{M}} \boldsymbol{U}_{\boldsymbol{M}}^\top$ to indicate the eigenvalue decomposition of matrix $\boldsymbol{M}$, the inverse of $\boldsymbol{A}$ can be rewritten as

$$\begin{aligned}
\boldsymbol{A}^{-1} &= \left[\left(\boldsymbol{U}_{\boldsymbol{C}_n} \boldsymbol{S}_{\boldsymbol{C}_n}^{1/2} \otimes \boldsymbol{I}_N \right) \left(\underbrace{\boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2} \boldsymbol{U}_{\boldsymbol{C}_n}^\top \boldsymbol{C}_g \boldsymbol{U}_{\boldsymbol{C}_n} \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2}}_{\boldsymbol{C}_g^\star} \otimes \boldsymbol{R}_g + \boldsymbol{I}_{NP} \right) \left(\boldsymbol{U}_{\boldsymbol{C}_n} \boldsymbol{S}_{\boldsymbol{C}_n}^{1/2} \otimes \boldsymbol{I}_N \right)^\top \right]^{-1} \\
&= \left[\left(\boldsymbol{U}_{\boldsymbol{C}_n} \boldsymbol{S}_{\boldsymbol{C}_n}^{1/2} \boldsymbol{U}_{\boldsymbol{C}_g^\star} \otimes \boldsymbol{U}_{\boldsymbol{R}_g} \right) \left(\boldsymbol{S}_{\boldsymbol{C}_g^\star} \otimes \boldsymbol{S}_{\boldsymbol{R}_g} + \boldsymbol{I}_{NP} \right) \left(\boldsymbol{U}_{\boldsymbol{C}_n} \boldsymbol{S}_{\boldsymbol{C}_n}^{1/2} \boldsymbol{U}_{\boldsymbol{C}_g^\star} \otimes \boldsymbol{U}_{\boldsymbol{R}_g} \right)^\top \right]^{-1} \\
&= \left(\boldsymbol{U}_{\boldsymbol{C}_g^\star}^\top \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2} \boldsymbol{U}_{\boldsymbol{C}_n}^\top \otimes \boldsymbol{U}_{\boldsymbol{R}_g}^\top \right)^\top \underbrace{\left(\boldsymbol{S}_{\boldsymbol{C}_g^\star} \otimes \boldsymbol{S}_{\boldsymbol{R}_g} + \boldsymbol{I}_{NP} \right)^{-1}}_{\boldsymbol{D}} \left(\underbrace{\boldsymbol{U}_{\boldsymbol{C}_g^\star}^\top \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2} \boldsymbol{U}_{\boldsymbol{C}_n}^\top}_{\boldsymbol{L}_c} \otimes \underbrace{\boldsymbol{U}_{\boldsymbol{R}_g}^\top}_{\boldsymbol{L}_r} \right) \\
&= \left(\boldsymbol{L}_c \otimes \boldsymbol{L}_r \right)^\top \boldsymbol{D} \underbrace{\left(\boldsymbol{L}_c \otimes \boldsymbol{L}_r \right)}_{\boldsymbol{L}},
\end{aligned} \tag{2.82}$$

where we introduced

$$\begin{aligned}
\boldsymbol{C}_g^\star &= \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2}\boldsymbol{U}_{\boldsymbol{C}_n}^\top \boldsymbol{C}_g \boldsymbol{U}_{\boldsymbol{C}_n}\boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2} && (2.83)\\
\boldsymbol{L}_c &= \boldsymbol{U}_{\boldsymbol{C}_g^\star}^\top \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2}\boldsymbol{U}_{\boldsymbol{C}_n}^\top && (2.84)\\
\boldsymbol{L}_r &= \boldsymbol{U}_{\boldsymbol{R}_g}^\top && (2.85)\\
\boldsymbol{L} &= \boldsymbol{L}_c \otimes \boldsymbol{L}_r && (2.86)\\
\boldsymbol{D} &= \left(\boldsymbol{S}_{\boldsymbol{C}_g^\star} \otimes \boldsymbol{S}_{\boldsymbol{R}_g} + \boldsymbol{I}_{NP}\right)^{-1}. && (2.87)
\end{aligned}$$

. Writing explicitly the log-marginal likelihood of the model and using (2.82)

$$\begin{aligned}
\log \mathcal{L} &= -NP\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{A} && (2.88)\\
&\quad -\frac{1}{2}\mathrm{vec}\left(\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{B}\right)^\top \boldsymbol{A}^{-1}\mathrm{vec}\left(\boldsymbol{Y}-\boldsymbol{F}\boldsymbol{B}\right) && (2.89)\\
&= -NP\log(2\pi) + \frac{1}{2}\log\det \boldsymbol{D} - \frac{N}{2}\log\det \boldsymbol{C}_n && (2.90)\\
&\quad -\frac{1}{2}\mathrm{vec}\left(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{F}}\tilde{\boldsymbol{B}}\right)^\top \boldsymbol{D}\,\mathrm{vec}\left(\tilde{\boldsymbol{Y}}-\tilde{\boldsymbol{F}}\tilde{\boldsymbol{B}}\right), && (2.91)
\end{aligned}$$

where we used

$$\log\det \boldsymbol{A} = -\log\det \boldsymbol{D} + N\log\det \boldsymbol{C}_n, \tag{2.92}$$

and introduced the transformed quantities

$$\begin{aligned}
\mathrm{vec}\left(\tilde{\boldsymbol{Y}}\right) &= \left(\boldsymbol{L}_c \otimes \boldsymbol{L}_r\right)\mathrm{vec}\left(\boldsymbol{Y}\right)\\
&= \mathrm{vec}\left(\boldsymbol{L}_r \boldsymbol{Y} \boldsymbol{L}_c^\top\right) && (2.93)\\
\tilde{\boldsymbol{F}} &= \boldsymbol{L}_r\boldsymbol{F} && (2.94)\\
\tilde{\boldsymbol{B}} &= \boldsymbol{B}\boldsymbol{L}_c^\top. && (2.95)
\end{aligned}$$

Note that, similarly to Section 2.3.3, we have introduced "transformed" phenotypes, which are linearly independent across all individuals and traits. This transformation is the Kronecker product between an across-individual (row) transformation $\boldsymbol{L}_r$ and an across-traits (column) transformation $\boldsymbol{L}_c$. The insight that $\boldsymbol{L}_r$ does not change during optimisation allows the derivation of an efficient algorithm, as row transformations needs be computed only once upfront. The full algorithm can be summarised in three steps:

1. Compute $\boldsymbol{L}_r$. This requires computation of the eigenvalue decomposition of the genetic relatedness matrix $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$, which has complexity $O(N^3)$.

2. Compute $\boldsymbol{L}_r\boldsymbol{Y}$ and $\boldsymbol{L}_r\boldsymbol{F}$, which has computational complexity $O(N^2P + N^2K)$.

3. Perform gradient based optimisation with the row-transformed quantities. This can be done in time complexity $O(N)$.

Section A.2 gives a full description of all computations for a more general version of this model.

In the next Chapter, we will generalise this efficient inference scheme to a multi-trait LMMs with two variance components.

Genome-wide testing

Genome-wide association testing requires maximum likelihood optimisation considering each of the genome-wide variants in turn. A strategy to reduce computation is to estimate variance components under the no-association model and then fix the structure of the total covariance in the genome-wide tests (Korte et al., 2012; Lippert et al., 2014c; Furlotte and Eskin, 2015). Briefly, introducing $\hat{\boldsymbol{A}} = \hat{\boldsymbol{C}}_g \otimes \boldsymbol{R} + \hat{\boldsymbol{C}}_n \otimes \boldsymbol{I}_N$, where $\hat{\boldsymbol{C}}_g$ and $\hat{\boldsymbol{C}}_n$ are the (restricted) MLE of $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ under the no-association model (Eq (2.79) with $\boldsymbol{b} = \boldsymbol{0}$), one can consider

$$\operatorname{vec}(\boldsymbol{Y}) \sim \mathcal{N}\left((\boldsymbol{I}_P \otimes \boldsymbol{g})\,\boldsymbol{b} + (\boldsymbol{I}_P \otimes \boldsymbol{F})\operatorname{vec}(\boldsymbol{B}),\ \sigma_t^2 \hat{\boldsymbol{A}}\right). \tag{2.96}$$

Similarly to the univariate approach discussed in Section 2.3.3, the MLE for the σ_t^2, $\boldsymbol{B}$ and $\boldsymbol{b}$ can be computed using close-form solutions. Overall, this strategy has a per-variant computational complexity of $O(N^2)$, which correspond to the transformation of the tested genotype vector, while other operations are linear in N.

We have here focused on studying how the model scales with the number of individuals. However, this approach does not not scale as well with the number of phenotypes (Zhou and Stephens, 2014), hindering applications to joint analysis across more than twenty traits. This is mainly due to the fact that the trait covariances are estimated by restricted maximum likelihood, which implies that (i) the number of parameters of the model increases quadratically with the number of traits and (ii) trait (column) computations cannot be cached as trait covariances change during optimisation. These concepts will be discussed in more detail in the next chapter, where I will discuss a generalization of the model considered here, enabling joint association testing of multiple variants and traits.

3 | Efficient set tests for joint analysis of correlated traits

As discussed in Section 2.3.5, set tests are statistical models that enable association testing between sets of genetic variants and a quantitative trait. Set tests can help reduce the number of genome-wide tests and have been shown to be more effective than single-variant approaches to detect effects due to either untyped variants (Wu et al., 2010) or multiple causal variants in linkage (Listgarten et al., 2013). Consequently, several set tests have been proposed for association testing with common (Wu et al., 2010; Listgarten et al., 2013) and rare variants (Wu et al., 2011; Lee et al., 2012c; Klaudia Walter et al., 2015). Within the linear mixed model (LMM) framework, the additive effects of the variants in a set can be aggregated using a variance component. However, LMMs that consider set tests for multiple traits are computationally demanding, and hence these methods have been limited to single-trait analysis.

The main hurdle for deriving an efficient multi-trait set test is that the speedups for single-variant testing rely on the fact that these LMMs only have one variance component (see Section 2.4.4). This is sufficient for single-variant association testing as the genetic tests are implemented using fixed effects. Conversely, LMM-based set tests require two variance components, one for modelling the joint effect of the variants in the set (set component) and a second component to account for genetic relatedness (relatedness component). In this chapter, I will present mtSet, a mixed-model approach that enables association testing between multiple variants and traits in a scalable manner while accounting for arbitrary confounding. mtSet builds on a matrix-variate LMM with two variance components, yet is computationally efficient and permits genetic analysis in large cohorts. An open-source software implementation of mtSet is available as part of the LIMIX software suite (http://github.com/PMBio/limix), which I discuss in Chapter 5.

In Section 3.1, I describe the mtSet model. In Sections 3.2, I discuss results from

experiments using simulated datasets, which were used to validate the model. Finally, in Section 3.3, I present results from two case studies from applying mtSet to real data.

The material presented in this chapter is joint work with Barbara Rakitsch, Christoph Lippert and Oliver Stegle, and it was published in *Nature Methods* in June 2015 (Casale et al., 2015).

3.1 A multi-trait set test

In this section, I introduce the mtSet model, discuss the strategy used to obtain P values and provide implementation details of the efficient algorithm. I also introduce faster versions of mtSet for analyses of cohorts with unrelated individuals and, finally, discuss the relationship of mtSet to existing methods.

3.1.1 The model

mtSet model is closely related to the matrix-variate mixed model introduced in Section 2.79. However, instead of a single variance component that accounts for genetic relatedness, we consider a second random effect that models the effect of the variants in the set. The matrix-variate phenotype $\boldsymbol{Y} \in \mathbb{R}^{N \times P}$ for N individuals and P traits is modelled as the sum of the contribution from fixed effect covariates, the variants in the genetic region (set component), a random genetic background effect (relatedness component) and residual observational noise

$$\boldsymbol{Y} = \underbrace{\boldsymbol{F}\boldsymbol{B}}_{\text{covariates}} + \underbrace{\boldsymbol{U}_r}_{\text{set component}} + \underbrace{\boldsymbol{U}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{\Psi}}_{\text{noise}}. \tag{3.1}$$

Here $\boldsymbol{F}$ is the $N \times K$ design matrix for K covariates, $\boldsymbol{B}$ the corresponding $K \times P$ weight matrix, $\boldsymbol{U}_r$ and $\boldsymbol{U}_g$ denote effects from the set and the relatedness components and $\boldsymbol{\Psi}$ is the residual noise. $\boldsymbol{U}_r$ and $\boldsymbol{U}_g$ and $\boldsymbol{\Psi}$ are modelled to follow matrix-variate normal distributions:

$$\begin{aligned} \boldsymbol{U}_r &\sim MVN(\boldsymbol{0}, \boldsymbol{R}_r, \boldsymbol{C}_r) \\ \boldsymbol{U}_g &\sim MVN(\boldsymbol{0}, \boldsymbol{R}_g, \boldsymbol{C}_g) \\ \boldsymbol{\Psi} &\sim MVN(\boldsymbol{0}.\boldsymbol{I}_N, \boldsymbol{C}_n), \end{aligned} \tag{3.2}$$

where $\boldsymbol{R}_r \in \mathbb{R}^{N \times N}$ and $\boldsymbol{R}_g \in \mathbb{R}^{N \times N}$ are the local and global realised relatedness matrices (RRMs), respectively. Specifically, denoting with $\boldsymbol{G} \in \mathbb{R}^{N \times R}$ the standardised

genotype matrix for the R variants in the set and with $\boldsymbol{S} \in \mathbb{R}^{N\times S}$ the standardised genotype matrix for S genome-wide variants, we define $\boldsymbol{R}_r = \frac{1}{R}\boldsymbol{G}\boldsymbol{G}^\top$ and $\boldsymbol{R}_g = \frac{1}{S}\boldsymbol{S}\boldsymbol{S}^\top$. $\boldsymbol{C}_r \in \mathbb{R}^{P\times P}$ and $\boldsymbol{C}_g \in \mathbb{R}^{P\times P}$ are the trait-to-trait covariance matrices of the set and the relatedness component, while $\boldsymbol{C}_n$ is the residual trait-to-trait covariance matrix.

Marginal likelihood. The marginal likelihood of the model in Eq (3.1-3.2) can be expressed as

$$\mathcal{L}(\boldsymbol{B},\boldsymbol{C}_r,\boldsymbol{C}_g,\boldsymbol{C}_n) = \mathcal{N}\left(\text{vec}(\boldsymbol{Y}) \middle| \text{vec}(\boldsymbol{F}\boldsymbol{B}),\ \underbrace{\boldsymbol{C}_r \otimes \boldsymbol{R}_r}_{\text{set component}} + \underbrace{\boldsymbol{C}_g \otimes \boldsymbol{R}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{C}_n \otimes \boldsymbol{I}_N}_{\text{noise}}\right) \quad (3.3)$$

where we used the Kronecker product and vec$(\cdot)$ operator (see Section 2.4.1). The case $P = 1$ corresponds to

$$\begin{aligned}
\boldsymbol{Y} \in \mathbb{R}^{N\times P} &\rightarrow \boldsymbol{y} \in \mathbb{R}^N,\\
\boldsymbol{B} \in \mathbb{R}^{K\times P} &\rightarrow \boldsymbol{b} \in \mathbb{R}^K,\\
\boldsymbol{C}_r \in \mathbb{R}^{P\times P} &\rightarrow \sigma_r^2 \in \mathbb{R},\\
\boldsymbol{C}_g \in \mathbb{R}^{P\times P} &\rightarrow \sigma_g^2 \in \mathbb{R},\\
\boldsymbol{C}_n \in \mathbb{R}^{P\times P} &\rightarrow \sigma_n^2 \in \mathbb{R}
\end{aligned}$$

and mtSet reduces to a single-trait set test

$$\mathcal{L}(\boldsymbol{b},\sigma_r^2,\sigma_g^2,\sigma_n^2) = \mathcal{N}\left(\boldsymbol{y} \middle| \boldsymbol{F}\boldsymbol{b},\ \underbrace{\sigma_r^2\boldsymbol{R}_r}_{\text{set component}} + \underbrace{\sigma_g^2\boldsymbol{R}_g}_{\text{relatedness component}} + \underbrace{\sigma_n^2\boldsymbol{I}_{N\times N}}_{\text{noise}}\right), \quad (3.4)$$

which we will denote with stSet. This special case of mtSet is equivalent to the model proposed in Lippert et al. (2014a).

Rank of the set component. As described in Section 3.1.3, a key insight to achieve computational efficiency is that the typical number of variants in the set is lower than the number of individuals ($R < N$), and thus the local RRM $\boldsymbol{R}_r = \frac{1}{R}\boldsymbol{G}\boldsymbol{G}^\top$ is low rank.

Denoting the rank of the trait-to-trait covariance matrix $\boldsymbol{C}_r$ with C, the total rank of the set component is

$$\text{rank}\,(\boldsymbol{C}_r \otimes \boldsymbol{R}_r) = CR. \quad (3.5)$$

Although mtSet allows considering any rank C of the trait covariance, we consider $C =$

1 in the experiments. Set trait covariances with higher ranks can also be considered, however at the cost of increased computational burden. As we will see in the next Chapter, considering higher-rank models allows for decomposition the local genetic effect into distinct genetic signals, where the rank C corresponds to the number of such signals.

Parametrisation of trait-to-trait covariances. In contrast to the individual-to-individual covariances, the three trait-to-trait covariance matrices $\boldsymbol{C}_r$, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ are estimated by maximising the restricted marginal likelihood.

To explicitly model the dependency of $\boldsymbol{C}_r$ on its rank C, we set $\boldsymbol{C}_r(\boldsymbol{E}) = \boldsymbol{E}\boldsymbol{E}^T$ where $\boldsymbol{E} \in \mathbb{R}^{P \times C}$ and consider the PC entries of $\boldsymbol{E}$ as model parameters. Conversely, for the relatedness and the noise trait-to-trait covariance matrices we consider general positive-definite matrices, which we parametrize in the Cholesky space. Specifically, we set $\boldsymbol{C}_g = \boldsymbol{L}_g\boldsymbol{L}_g^T$ and $\boldsymbol{C}_n = \boldsymbol{L}_n\boldsymbol{L}_n^T$ where $\boldsymbol{L}_g$ and $\boldsymbol{L}_n$ are lower triangular matrices and consider the non-zero entries of $\boldsymbol{L}_g$ and $\boldsymbol{L}_n$ as model parameters. In this way, both $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ are parametrized by $\frac{1}{2}P(P+1)$, i.e. the number of non-zeros entries of a lower diagonal matrix.

Maximisation of the marginal likelihood. To fit the model, we optimise the log restricted maximum likelihood with respect to the weight matrix $\boldsymbol{B}$ and the three trait-to-trait covariances $\boldsymbol{C}_r$, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$, using a gradient-based parameter optimiser. In our implementation, we use a low-memory BFGS (Liu and Nocedal, 1989; Zhu et al., 1997) as implemented in the *fmin_l_bfgs_b* optimisation method in the SciPy python library (Jones et al., 2001). To improve convergence for datasets with smaller sample sizes ($N < 5,000$) and large windows we increase accuracy of the optimisation procedure by setting the SciPy default parameter *factr* to 10^3 (the SciPy default value is 10^7). As shown in Section 3.1.3, a combination of low-rank covariance updates, eigenvalue decomposition-based speedups and Kronecker product properties allows achieving efficient evaluations of the likelihood function and its gradients. Prior to model fitting, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ are initialised to their maximum likelihood estimator (MLE) for $\boldsymbol{C}_r = \boldsymbol{0}$ (i.e., from the model in Eq (2.78)). This strategy exploits that (i) learning $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ is very efficient (see Section 2.4.4) and (ii) in many settings variant-sets do not explain a high proportion of the phenotypic variance so that the MLE of the full model is very close to this starting point.

Choice of variant sets. Different strategies can be adopted to define the variant-sets to be tested. Common choices include gene-set analyses (Wu et al., 2010; Listgarten et al., 2013) and sliding window approaches (Wu et al., 2011). In the experiments presented here we opted for the latter, as depending on the window size a gene-based approach can result in an uneven representation of the genome, with high gene-density regions corresponding to a large number of overlapping sets and low gene-density regions corresponding to a few or no sets. However, gene-based analyses can be implemented in the same vein and are supported by the mtSet implementation. Another important parameter to choose is the size of the variant sets. In Section 3.2.5 we use simulated data to study how the choice of the window size affects the statistical power and the computational efficiency of the method.

3.1.2 Statistical testing

In mtSet, we test for association between the considered set of variants and any of the modelled traits by assessing whether the set component is significantly different from zero. More formally, we compare the null model without the set component

$$\mathcal{M}_0 : \mathrm{vec}\,(\boldsymbol{Y}) \sim \mathcal{N}\left(\mathrm{vec}(\boldsymbol{Y}) \middle| \mathrm{vec}(\boldsymbol{FB}),\ \underbrace{\boldsymbol{C}_g \otimes \boldsymbol{R}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{C}_n \otimes \boldsymbol{I}_{N\times N}}_{\text{noise}}\right) \tag{3.6}$$

with the alternative model

$$\mathcal{M}_1 : \mathrm{vec}\,(\boldsymbol{Y}) \sim \mathcal{N}\left(\mathrm{vec}(\boldsymbol{Y}) \middle| \mathrm{vec}(\boldsymbol{FB}),\ \underbrace{\boldsymbol{C}_r \otimes \boldsymbol{R}_r}_{\text{set component}} + \underbrace{\boldsymbol{C}_g \otimes \boldsymbol{R}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{C}_n \otimes \boldsymbol{I}_{N\times N}}_{\text{noise}}\right) \tag{3.7}$$

using the log likelihood ratio (LLR) test statistics (see Section 2.2.1). As for variance component tests the parameter space is constrained and the null hypothesis lies on the boundaries of the parameter space (Molenberghs and Verbeke, 2007; Self and Liang, 1987; Dominicus et al., 2006) Wilk's theorem does not apply, meaning that the asymptotic distribution of the LLR test statistics is generally not known. Indeed, the trait-to-trait covariance matrices are constrained to be positive-semidefinite ($\boldsymbol{C}_r, \boldsymbol{C}_g, \boldsymbol{C}_n \succeq 0$) and the null hypothesis is $\boldsymbol{C}_r = \boldsymbol{0}$.

For single-trait set tests ($P = 1$), the constraint corresponds to the non-negativity condition of the set variance component ($\sigma_r^2 \geq 0$ in Eq (3.4)) while the null model corresponds to $\sigma_r^2 = 0$. Self and Liang (1987) showed that, under certain regulatory conditions, the asymptotic distribution of the LLR statistics is a 50-50 mixture of two

χ^2 components, the first with 0 degrees of freedom (dof) and the second with one dof

$$2LLR \sim 0.5\chi^2_0 + 0.5\chi^2_1. \tag{3.8}$$

Intuitively the first χ^2_0 component (which is equivalent to a Dirac function $\delta(x)$) captures the cases where the constrained MLE $\hat{\sigma}^2_r$ of σ^2_r is at the bound ($\hat{\sigma}^2_r = 0$) while the second term describes cases where $\hat{\sigma}^2_r > 0$. As suggested by the coefficients of the mixture, the two scenarios are expected to occur with probability 50%. However, one of the necessary conditions for this asymptotic form is that multiple sub-vectors of the output variable are identically distributed, which is unlikely to hold in genetic analyses. In practice this results in conservative P values as the realised mixing weight of the χ^2_0 component is higher than 0.5 (Greven et al., 2012; Listgarten et al., 2013). To overcome this Listgarten et al. (2013) proposed relaxing the distributions to

$$\frac{1}{a}LLR \sim \pi\chi^2_0 + (1-\pi)\chi^2_d, \tag{3.9}$$

where a, π and d are fit to match the empirical distribution of null tests obtained from permutations. We have adopted this strategy within mtSet and empirically validated this approach for multi-trait set tests.

Obtaining P values in mtSet. Let us assume we have performed a set-based GWAS consisting of T genome-wide set tests. Following Listgarten et al. (2013), we employ the following three-step procedure to compute P values:

1. **Obtain the empirical distribution using permutations.** We consider J permutations of individuals in the set components and for each permutation we perform a genome-wide scan (consisting of T set tests). The empirical distribution of the LLR under $\mathcal{H}_0$ is then obtained by pooling the JT test statistics across all tests and permutations[1]. Note that this permutation strategy i) keeps intact the LD structure and MAF distribution of real genotypes and ii) retains the dependency between the relatedness component, the fixed effect covariates and the phenotypes.

2. **Fit the parametric form to the empirical distribution.** We consider the parametric form in Eq. (3.9). The P value corresponding to a LLR test statistics

[1]By pooling across all tests, we assume that the distribution of the LLRs under the null model is the same across all sets.

D can be obtained as

$$\mathrm{P}(D;\pi,a,d) = (1-\pi)\left(1 - F\left(\frac{D}{a},d\right)\right), \tag{3.10}$$

where $F(\cdot,d)$ is the cumulative density function of $\chi^2_d(\cdot)$. The parameters π, a and d are empirically determined by fitting the parametric form to the obtained distribution in a two-step procedure. First, we estimate $\hat{\pi}$ as the fraction of LLRs that are lower than a tolerance value (which we set to $3 \cdot 10^{-4}$). Second, we determine a and d by minimising the squared error between $\log \mathrm{P}(D;\hat{\pi},a,d)$ and the log of the corresponding theoretical values under $\mathcal{H}_0$ (P values are uniformly distributed $\mathcal{H}_0$). To prioritise accuracy for low P values, we only consider the lowest 10% of the P values to determine a and d. Sorting the T LLRs from the largest to the smallest, i.e., $D_1 \geq D_2 \geq \cdots \geq D_T$, and indicating with I the number of LLRs in the last decile, the mean squared error function we consider is

$$\mathrm{MSE}(a,d) = \frac{1}{I}\sum_{i=1}^{I}\left(\log \mathrm{P}(D;\hat{\pi},a,d) - \log\left(\frac{i}{JT+1}\right)\right)^2. \tag{3.11}$$

The MSE is minimised by grid search with $a \in [0.1, 5.0]$, $d \in [0.1, 5.0]$ and considering 100 equally spaced values for both parameters. Although only the top 10% of the null test statistics are used in the fit of the parametric distribution, we find empirically a good fit for the complete range of the test statistics. **Fig. B.1** shows five examples of fits from the calibration experiment using genotypes from the 1000 Genome Project (see Section 3.2.4). The estimated parameters for the two real-data analyses described in Section 3.3 are reported in **Table B.1**.

3. **Pv estimation.** Once the optimal parameter values $\hat{\pi}$, $\hat{a}$ and $\hat{d}$ are determined, the P value can be calculated from the test statistics as $\mathrm{P}(D;\hat{\pi},\hat{a},\hat{d})$.

Given the large number of genome-wide tests, we found that approximately 30 genome-wide permutations are sufficient to estimate the null distribution and obtain calibrated P values in our experiments.

3.1.3 Efficient parameter inference

In this section, I outline the efficient algebraic computation of the log marginal likelihood (LML) implemented in mtSet. Derivations for efficient computation of the gradients of mtSet are given in Section A.3.

Log-marginal likelihood. The log likelihood of the model in Eq (3.3) is given by

$$\mathcal{L} = -NP\log 2\pi - \frac{1}{2}\log\det \boldsymbol{K} - \frac{1}{2}\mathrm{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right), \tag{3.12}$$

where

$$\begin{aligned}
\boldsymbol{K} &= \boldsymbol{C}_r \otimes \boldsymbol{R}_r + \boldsymbol{C}_g \otimes \boldsymbol{R}_g + \boldsymbol{C}_n \otimes \boldsymbol{I}_N && (3.13)\\
&= \frac{1}{R}\boldsymbol{E}\boldsymbol{E}^\top \otimes \boldsymbol{G}\boldsymbol{G}^\top + \boldsymbol{C}_g \otimes \boldsymbol{R}_g + \boldsymbol{C}_n \otimes \boldsymbol{I}_N && (3.14)\\
&= \underbrace{\left(\frac{1}{\sqrt{R}}\boldsymbol{E} \otimes \boldsymbol{G}\right)}_{\boldsymbol{X}} \left(\frac{1}{\sqrt{R}}\boldsymbol{E} \otimes \boldsymbol{G}\right)^\top + \underbrace{\boldsymbol{C}_g \otimes \boldsymbol{R}_g + \boldsymbol{C}_n \otimes \boldsymbol{I}_N}_{\boldsymbol{A}} && (3.15)\\
&= \boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{A}. && (3.16)
\end{aligned}$$

In Eq (3.16) we have made the low rank structure of the set component explicit, by writing $\boldsymbol{C}_r = \boldsymbol{E}\boldsymbol{E}^T$ and $\boldsymbol{R}_r = \frac{1}{R}\boldsymbol{G}\boldsymbol{G}^T$, and we have introduced

$$\begin{aligned}
\boldsymbol{X} &= \frac{1}{\sqrt{R}}\boldsymbol{E} \otimes \boldsymbol{G} && (3.17)\\
\boldsymbol{A} &= \boldsymbol{C}_g \otimes \boldsymbol{R}_g + \boldsymbol{C}_n \otimes \boldsymbol{I}_N. && (3.18)
\end{aligned}$$

Note that $\boldsymbol{A}$ is the total covariance matrix of the model in Eq (2.78) that is used for single-variant association testing across multiple traits.

The bottleneck operations in the computation of the likelihood in Eq (3.12) are $\boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right)$ and the log determinant of $\boldsymbol{K}$. However, in the following we will see how these calculations can be performed efficiently by exploiting the low-rank nature of $\boldsymbol{X}\boldsymbol{X}^\top$.

Rewriting the inverse. Using the Woodbury identity (Woodbury, 1950) we can re-write the inverse of $\boldsymbol{K}$ as

$$\begin{aligned}
\boldsymbol{K}^{-1} &= \left(\boldsymbol{A} + \boldsymbol{X}\boldsymbol{X}^\top\right)^{-1}\\
&= \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{X}\left(\boldsymbol{I} + \boldsymbol{X}^\top\boldsymbol{A}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{A}^{-1} && (3.19)
\end{aligned}$$

Using Eq (2.82), which shows that

$$\boldsymbol{A}^{-1} = \boldsymbol{L}^T\boldsymbol{D}\boldsymbol{L} \tag{3.20}$$

with

$$\begin{aligned}
\boldsymbol{D} &= \left(\boldsymbol{S}_{\boldsymbol{C}_g^\star} \otimes \boldsymbol{S}_{\boldsymbol{R}_g} + \boldsymbol{I}\right)^{-1} && (3.21)\\
\boldsymbol{L} &= \boldsymbol{L}_c \otimes \boldsymbol{L}_r && (3.22)\\
\boldsymbol{L}_r &= \boldsymbol{U}_{\boldsymbol{R}_g}^T && (3.23)\\
\boldsymbol{L}_c &= \boldsymbol{U}_{\boldsymbol{C}_g^\star}^T \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2} \boldsymbol{U}_{\boldsymbol{C}_n}^T && (3.24)\\
\boldsymbol{C}_g^\star &= \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2} \boldsymbol{U}_{\boldsymbol{C}_n}^T \boldsymbol{C}_g \boldsymbol{U}_{\boldsymbol{C}_n} \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2}, && (3.25)
\end{aligned}$$

we have

$$\begin{aligned}
\boldsymbol{K}^{-1} &= \boldsymbol{L}^T\boldsymbol{D}\boldsymbol{L} - \boldsymbol{L}^T\boldsymbol{D}\underbrace{\boldsymbol{L}\boldsymbol{X}}_{\boldsymbol{W}}\left(\boldsymbol{I} + \boldsymbol{X}^\top\boldsymbol{L}^\top\boldsymbol{D}\boldsymbol{L}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{L}^T\boldsymbol{D}\boldsymbol{L}\\
&= \boldsymbol{L}^\top\boldsymbol{D}\boldsymbol{L} - \boldsymbol{L}^\top\boldsymbol{D}\boldsymbol{W}\left(\underbrace{\boldsymbol{I} + \boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{W}}_{\boldsymbol{\Lambda}}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{L}\\
&= \boldsymbol{L}^\top\boldsymbol{D}\boldsymbol{L} - \boldsymbol{L}^\top\boldsymbol{D}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{L}. && (3.26)
\end{aligned}$$

where we introduced

$$\begin{aligned}
\boldsymbol{\Lambda} &= \boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W} \in \mathbb{R}^{RC\times RC} && (3.27)\\
\boldsymbol{W} &= \boldsymbol{L}\boldsymbol{X}\\
&= \underbrace{\boldsymbol{L}_c\boldsymbol{E}}_{\boldsymbol{W}_c} \otimes \underbrace{\frac{1}{\sqrt{R}}\boldsymbol{L}_r\boldsymbol{G}}_{\boldsymbol{W}_c}\\
&= \boldsymbol{W}_c \otimes \boldsymbol{W}_r && (3.28)\\
\boldsymbol{W}_c &= \boldsymbol{L}_c\boldsymbol{E} \in \mathbb{R}^{P\times C} && (3.29)\\
\boldsymbol{W}_r &= \frac{1}{\sqrt{R}}\boldsymbol{L}_r\boldsymbol{G} \in \mathbb{R}^{N\times R}. && (3.30)
\end{aligned}$$

Computation of the column matrix $\boldsymbol{W}_c$ has $O(P^2C)$ complexity, while computation of the row covariance matrix $\boldsymbol{W}_r$ has $O(N^2R)$ computational complexity. Note that $\boldsymbol{W}_r$ is constant during optimisation and thus it needs to be computed only once per region.

The matrix $\boldsymbol{\Lambda}$ can also be computed efficiently by rewriting it as

$$\begin{aligned}
\boldsymbol{\Lambda} &= \boldsymbol{I} + \boldsymbol{W}^T \boldsymbol{D}\boldsymbol{W} && (3.31)\\
&= \boldsymbol{I} + \left(\boldsymbol{W}_c^T \otimes \boldsymbol{W}_r^T\right)\left[\boldsymbol{D}\boldsymbol{W}_{:,1} \quad \ldots \quad \boldsymbol{D}\boldsymbol{W}_{:,RC}\right] && (3.32)\\
&= \boldsymbol{I} + \left[\mathrm{vec}\left(\boldsymbol{W}_r^T\left(\mathrm{vec}^{-1}\left(\boldsymbol{D}\boldsymbol{W}_{:,1}\right)\boldsymbol{W}_c\right)\right) \quad \ldots \right. && (3.33)\\
&\qquad \left. \ldots \quad \mathrm{vec}\left(\boldsymbol{W}_r^T\left(\mathrm{vec}^{-1}\left(\boldsymbol{D}\boldsymbol{W}_{:,RC}\right)\boldsymbol{W}_c\right)\right)\right]. && (3.34)
\end{aligned}$$

Indeed, computing $\boldsymbol{W}$ explicitly and computing $\boldsymbol{D}\boldsymbol{W}$ takes $O(CNPR)$ time, multiplying $\boldsymbol{D}\boldsymbol{W}$ by $\boldsymbol{W}$ from left has complexity $O(CR(NPC + RNC))$, while the inversion of $\boldsymbol{\Lambda}$ has complexity $O(C^3R^3)$. In practice, we use the Cholesky factorization to compute the inverse of $\boldsymbol{\Lambda}$ having the advantage that we can re-use the decomposition for computing the log determinant later on.

Evaluating the log-determinant. The log-determinant can be computed by using the matrix determinant lemma (Harville, 1998)

$$\begin{aligned}
\log\det \boldsymbol{K} &= \log\det\left(\boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{A}\right) && (3.35)\\
&= \log\det \boldsymbol{A} + \log\det(\underbrace{\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}}_{\boldsymbol{\Lambda}}). && (3.36)
\end{aligned}$$

Provided that we have already computed $\boldsymbol{L}_c, \boldsymbol{L}_r, \boldsymbol{D}$ and the Cholesky decomposition of $\boldsymbol{\Lambda}$, evaluating $\log\det \boldsymbol{A}$ takes $O(NT)$ using Eq (2.92) while evaluating $\log\det\boldsymbol{\Lambda}$ takes $O(CR)$.

Evaluating the squared form. The squared form can be evaluated as follows

$$\begin{aligned}
\mathrm{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right) &= \mathrm{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{L}^T\boldsymbol{D}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right)\\
&\quad -\mathrm{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{L}^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}\underbrace{\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right)}_{\mathrm{vec}\bar{\boldsymbol{Y}}}\\
&= \mathrm{vec}\left(\tilde{\boldsymbol{Y}}\right)^T \boldsymbol{D}\ \mathrm{vec}\left(\tilde{\boldsymbol{Y}}\right) - \mathrm{vec}\left(\bar{\boldsymbol{Y}}\right)^T \boldsymbol{\Lambda}^{-1}\mathrm{vec}\left(\bar{\boldsymbol{Y}}\right),
\end{aligned}$$

where we have defined

$$\begin{aligned} \text{vec}\left(\tilde{\boldsymbol{Y}}\right) &= \boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right) \\ &= \text{vec}\left(\boldsymbol{L}_r \boldsymbol{Y} \boldsymbol{L}_c^T\right) \qquad (3.37) \\ \text{vec}\left(\bar{\boldsymbol{Y}}\right) &= \boldsymbol{W}^T \boldsymbol{D}\text{vec}\left(\tilde{\boldsymbol{Y}}\right) \\ &= \text{vec}\left(\boldsymbol{W}_r^T \boldsymbol{D}\text{vec}\left(\tilde{\boldsymbol{Y}}\right) \boldsymbol{W}_c\right). \qquad (3.38) \end{aligned}$$

Rotating and scaling the data Eqs (3.37 - 3.38) takes $O(N^2P + NP^2 + NP + NPC + RNC)$ time, where again the $O(N^2P)$ operation is done only once prior to the analysis. Computing the squared form $\text{vec}\left(\bar{\boldsymbol{Y}}\right)^T \boldsymbol{\Lambda}^{-1}\text{vec}\left(\bar{\boldsymbol{Y}}\right)$ takes $O(C^2R^2)$ time after having inverted $\boldsymbol{\Lambda}$.

The efficient derivation of gradients is described in Section A.3. Overall, combining eigenvalue decomposition and low-rank updates with Kronecker product algebra the $O(N^3P^3)$ computational cost can be reduced to a $O(N^3)$ operation upfront and $O(N^2 + NR^2P^2 + NRP^4)$ per set. Note that the quadratic operation in N needs to be performed only once for each set while gradient-based optimisation is linear in N (see **Table B.2**).

3.1.4 Analyses of cohorts with unrelated individuals

As any exact LMM, mtSet is bound to the upfront eigenvalue decomposition of the genetic relatedness matrix, which has complexity $O(N^3)$, limiting the scalability of LMMs to very large cohorts ($N \geq 20,000$). To overcome this, we considered two different strategies that can be employed for analysis of unrelated individuals: mtSet-PC and mtSet-LowRankBg. Both models scale linearly with the number of samples, enabling analysis of cohorts with up to 500,000 unrelated individuals (**Fig. 3.1**). In mtSet-PC, the relatedness component is dropped and population structure is accounted for by modelling the top principal components of the RRM as fixed effect covariates. In mtSet-LowRankBg, we instead consider a low-rank approximation of the RRM. The two methods build on the same modelling assumption, i.e., they both use a low-rank representation of the RRM, and gave similar results in simulated data (see **Fig. B.2**). For this reason we considered only mtSet-PC in simulations and real-data experiments. Full details on the algebraic implementation of the two methods are provided in Sections A.4-A.5.

3.1.5 Relationship to existing methods

mtSet is related to the single-variant LMM for multiple traits (Section 2.4) and to single-trait set tests (Section 2.3.5). In the following, I briefly review the computational complexity of related LMM implementations.

LMMs with fixed effect testing. As discussed in detail in Section 2.3.3, single-variant LMMs with fixed effect testing require a single $O(N^3)$ operation up-front and a per-test complexity of $O(N^2)$. Considering a RRM with low-rank structure (for example using feature selection, see Listgarten et al. (2012)), the computational complexity of these approaches can be further reduced to $O(N^2)$ for the up-front computation and a per-test complexity of $O(N)$. The multi-trait extension of this model (see Section 2.4.4) requires a single $O(N^3)$ operation up-front and $O(N^2+NP^x)$ per variant, where x depends on the optimisation algorithm. In particular, the efficient implementation proposed by Zhou and Stephens (mvLMM 2014) combines Newton-Raphson (Thompson, 1973) and the the PX-EM (Liu et al., 1998; Foulley and Van Dyk, 2000) algorithms.

LMMs for set tests. In the same way mvLMM extends a standard single-variant LMM to multi-trait analysis, mtSet can be regarded as the multivariate generalisation of the full-rank fast LMM-set proposed in Lippert et al. (2014a). The computational cost of mtSet consists in $O(N^3)$ operations upfront and $O(N^2 + NR^2P^2 + NRP^4)$ per set, where R denotes the number of variants in the set. We also considered two alternative low-rank approximations of the full mtSet model: mtSet-PC, where the relatedness component is omitted and population structure is modelled as a fixed effect, and mtSet-LowRankBg, which assumes a low-rank RRM. mtSet-LowRankBg can be regarded as the multi-trait extension of FaST-LMM-Set (Listgarten et al., 2013), which uses the same strategy to achieve fast computations in single-trait analyses.

Table B.3 provides a tabular listing of the per-test computational complexity for alternative LMM methods and implementations. Note that the listed complexities do not take into account the upfront $O(N^3)$ operation[2] for the eigen decomposition of the relatedness covariance matrix that.

[2]Or $O(N^2)$ if a low rank relatedness covariance is used

3.2 Simulation study

First, we validated mtSet and assessed its scalability, calibration and statistical power using simulations. Genetic effects were simulated based on genotype data from the 1000 Genomes project (Phase 1, 1000 Genomes Project Consortium (2012)). In Sections 3.2.1-3.2.2 I describe the simulation strategy, in Sections 3.2.3-3.2.5 I discuss the results.

3.2.1 Genotype simulation strategy

To study scalability and calibration of mtSet, we generated cohorts with different sample sizes and types of confounding based on genotype data from 1000 Genomes Project (Phase 1), consisting of 1,092 individuals and 9,034,769 genome-wide variants with a minor allele frequency of at least 2%. To do so, we followed the two-step procedure proposed in Loh et al. (2015b):

1. for each newly synthesised individual we randomly select A ancestors;

2. the new genotype is built as a mosaic of blocks of 1,000 variants, where each block is copied from the corresponding block in one of the A ancestors (selected at random).

As discussed in Loh et al. (2015b), the amount of genetic relatedness in the cohort depends on the number of ancestors A. For example, $A = 10$ gives rise to a cohort with approximately unrelated individuals while $A = 2$ gives a cohort with some highly related individuals, thereby simulating family relatedness. Moreover, by sampling all the ancestors of an individual either from the same sub-population or randomly from the whole cohort, one can generate genotype data with or without population structure, respectively.

Genotype data for runtime experiments. For runtime experiments, we generated cohorts with 500, 1K, 2K, 5K, 10K, 20K, 50K, 100K, 200K and 500K individuals. We first assigned each individual to one of the 14 populations in the 1000 Genomes Project and then we sampled $A = 10$ ancestors from that population, generating cohorts of unrelated individuals and preserving the population structure in the original data.

Genotype data for calibration experiments. For calibration experiments, we generated three cohorts with 1,000 individuals and different genetic structures, considering only European populations (CEU, FIN, GBR, IBS and TSI).

- **Population Structure** (*simPopStructure*). We simulated population structure by (i) assigning each newly synthesised individual to one of the five European populations (CEU, FIN, GBR, IBS and TSI) and (ii) considering $A = 10$ ancestors from that population;

- **Unrelated individuals** (*simUnrelated*). For each synthesised individual, we considered $A = 10$ ancestors, which were randomly sampled from the whole set of Europeans;

- **Related Individuals** (*simRelated*). For each synthesised individual, we considered $A = 2$ ancestors, again sampled from the whole set of Europeans.

The RRM for the original 1000 Genomes project data and the three synthetic cohorts is shown in **Fig. B.3**.

3.2.2 Phenotype simulation strategy

Quantitative traits were simulated assuming a linear-additive model, considering the contributions from a randomly selected causal genetic region for the set component, polygenic background effects from all remaining genome-wide variants, a contribution from unobserved covariates and iid observational noise. When assessing calibration of significance levels, phenotypes were generated from the null model, omitting the effect from the set component. To study the performance of the proposed set tests under different genetic architectures, we varied the variance explained by the selected region (h_r^2), the number of causal variants in the region (S_r), the percentage of shared causal variants (π_r), the variance explained by relatedness effects (h_g^2), the fraction of residual variance that is not iid across samples (λ), and the fraction of relatedness and residual signal that is shared across traits (α). Unless specified otherwise, we consider a region of 30 kb and use default values of the simulation parameters in **Table B.4**, which were chosen to mimic trait architectures observed in the rat data and the NFBC study (see Sections 3.3.1-3.3.2). In the following I provide details on the simulation pipeline and give a precise definition of the simulation parameters.

1. *Set contribution.*

$$
\begin{aligned}
S_r &= \text{average number of causal variants from the region across traits,} \\
h_r^2 &= \text{average fraction of variance explained by causal variants across traits,} \\
\pi_r &= \text{fraction of causal variants that have shared signal across traits}
\end{aligned}
$$

- **Set shared signal**
  We first select $S^{(\mathrm{s})} = \pi_r S_r$ variants. Then we sample a binary vector $\boldsymbol{b}^{(\mathrm{s})}_{\mathrm{snps}} \in \{-1, 1\}^{S^{(\mathrm{s})}}$ that defines the relative directionality of the effects across variants and a binary vector $\boldsymbol{b}^{(\mathrm{s})}_{\mathrm{traits}} \in \{-1, 1\}^{P}$ that defines the relative directionality of the effects across traits. Defining the $S^{(\mathrm{s})} \times P$ weight matrix as $\boldsymbol{B}^{(\mathrm{s})} = \boldsymbol{b}^{(\mathrm{s})}_{\mathrm{snps}} \otimes \boldsymbol{b}^{(\mathrm{s})\,T}_{\mathrm{traits}}$ and indicating with $\boldsymbol{X}^{(\mathrm{s})} \in \mathbb{R}^{N \times S^{(\mathrm{s})}}$ the standardised genotypes of the $S^{(\mathrm{s})}$ variants, the shared signal $\boldsymbol{R}^{(\mathrm{s})}$ from the set component is defined as
  $$\boldsymbol{R}^{(\mathrm{s})} = \boldsymbol{X}^{(\mathrm{s})} \boldsymbol{B}^{(\mathrm{s})} \tag{3.39}$$

- **Set independent signal**
  For each trait we select $S_r - S^{(\mathrm{s})}$ variants on average across traits and generate their effect sizes as iid from $\{-1, 1\}$. Indicating with $\boldsymbol{X}^{(\mathrm{i})}_p \in \mathbb{R}^{N \times S^{(\mathrm{i})}_p}$ the standardised genotype matrix for the $S^{(\mathrm{i})}_p$ causal variants selected for trait p and with $\boldsymbol{b}^{(\mathrm{i})}_p \in \{-1, 1\}^{S^{(\mathrm{i})}_p}$ their effect sizes on trait p, the independent signal is defined as
  $$\boldsymbol{R}^{(\mathrm{i})} = \begin{bmatrix} \boldsymbol{X}^{(\mathrm{i})}_1 \boldsymbol{b}^{(\mathrm{i})}_1 & \cdots & \boldsymbol{X}^{(\mathrm{i})}_P \boldsymbol{b}^{(\mathrm{i})}_P \end{bmatrix} \tag{3.40}$$

$\boldsymbol{R} = \boldsymbol{R}^{\mathrm{comm}} + \boldsymbol{R}^{\mathrm{ind}}$ is then scaled such that the average variance of its columns is h^2_r.

2. *Relatedness contribution.*

   h^2_g = average fraction of variance explained by relatedness across traits,
   α_g = fraction of signal from the relatedness contribution that is shared across traits

   - **Relatedness shared signal**
     The shared signal from relatedness is generated as
     $$\boldsymbol{G}^{(\mathrm{s})} \sim \mathrm{MVN}\left(\boldsymbol{0},\, \boldsymbol{R}_g,\, \boldsymbol{a}\boldsymbol{a}^T\right), \quad \text{with } \boldsymbol{a} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_P\right) \tag{3.41}$$
     and rescaled such that the average variance of its columns is αh^2_g.

   - **Relatedness independent signal**
     The independent signal from relatedness is generated as
     $$\boldsymbol{G}^{(\mathrm{i})} \sim \mathrm{MVN}\left(\boldsymbol{0},\, \boldsymbol{R}_g,\, \mathrm{diag}\left(\boldsymbol{c}^2\right)\right), \quad \text{with } \boldsymbol{c} \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_P\right) \tag{3.42}$$

and rescaled such that the average variance of its columns is $(1-\alpha)h_g^2$.

3. *Residuals.*

$$\begin{aligned} \lambda &= \text{fraction of structured noise,} \\ \alpha_H &= \text{fraction of structured noise that is shared across traits} \end{aligned}$$

- **Structured noise**
  We generate the input matrix for $K = 10$ unobserved covariates $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times K}$ as iid from $\mathcal{N}(0,1)$.

  - **Structured-noise shared signal**
    The shared signal from structured noise is generated as

$$\boldsymbol{H}^{(\mathrm{s})} \sim \mathrm{MVN}\left(\mathbf{0},\, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top,\, \boldsymbol{a}_H\boldsymbol{a}_H^\top\right), \quad \text{with } \boldsymbol{a}_H \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_P) \tag{3.43}$$

    and rescaled such that the average variance of its columns is $\alpha_H\lambda(1 - h_r^2 - h_g^2)$.

  - **Structured-noise independent signal**
    The independent signal from relatedness is generated as

$$\boldsymbol{H}^{(\mathrm{i})} \sim \mathrm{MVN}\left(\mathbf{0},\, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top,\, \mathrm{diag}\left(\boldsymbol{c}_H^2\right)\right), \quad \text{with } \boldsymbol{c}_H \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_P) \tag{3.44}$$

    and rescaled such that the average variance of its columns is $(1 - \alpha_H)\lambda(1 - h_r^2 - h_g^2)$.

- **IID noise**
  We generate iid Gaussian noise $\boldsymbol{E} \sim \prod \mathcal{N}(0,1)$ that we rescale so that the average variance of its columns is $(1-\lambda)(1 - h_r^2 - h_g^2)$.

We considered $\alpha_g = \alpha_H = \alpha$ in all simulation experiments. Overall, the phenotype can be expressed as

$$\boldsymbol{Y} = \boldsymbol{R}^{(\mathrm{s})} + \boldsymbol{R}^{(\mathrm{i})} + \boldsymbol{G}^{(\mathrm{s})} + \boldsymbol{G}^{(\mathrm{i})} + \boldsymbol{H}^{(\mathrm{s})} + \boldsymbol{H}^{(\mathrm{i})} + \boldsymbol{E}, \tag{3.45}$$

where the average variance across traits of each term can be written as a function of the simulation parameters introduced as shown in **Table B.5**.

3.2.3 Empirical complexity and scalability

To assess the scalability of mtSet, we considered synthetic cohorts with increasing sample sizes (up to 500,000 individuals), which were generated as described in Section 3.2.1. Phenotype data were generated as described in Section 3.2.2. However, to avoid computation of the RRM and its Cholesky decomposition for big cohorts ($N > 20,000$) we used a low-rank approximation of the RRM based on population labels when simulating phenotypes. Specifically, denoting with Π the number of populations and $\boldsymbol{J} \in \mathbb{R}^{N \times \Pi}$ an indicator matrix such that $\boldsymbol{J}_{ij}$ is 1 if individual i is from population j and 0 otherwise, we considered $\frac{1}{\Pi}\boldsymbol{J}\boldsymbol{J}^\top$ as relatedness matrix. We then assessed empirical runtime to complete a sliding-window experiment with a window size of 30kb and a step of 15kb on chromosome 20 (which resulted in 3,975 analysed regions) for the following set test implementations:

1. A naive implementation of the full mtSet model (without efficient algebra implementation, mtSet-naive);

2. The full mtSet model, which uses a random effect to model genetic relatedness (mtSet);

3. mtSet-PC, including the population labels $\boldsymbol{J}$ as fixed effect covariates to account for population structure;

4. mtSet-LowRankBg, considering a low-rank relatedness matrix $\frac{1}{\Pi}\boldsymbol{J}\boldsymbol{J}^\top$.

Empirical runtimes were assessed as the CPU time on a single core of an Intel Xeon CPU E5-2670 2.60 GHz processor. The reported runtime includes the cost for fitting the null model (only once) and the alternative models (for each test).

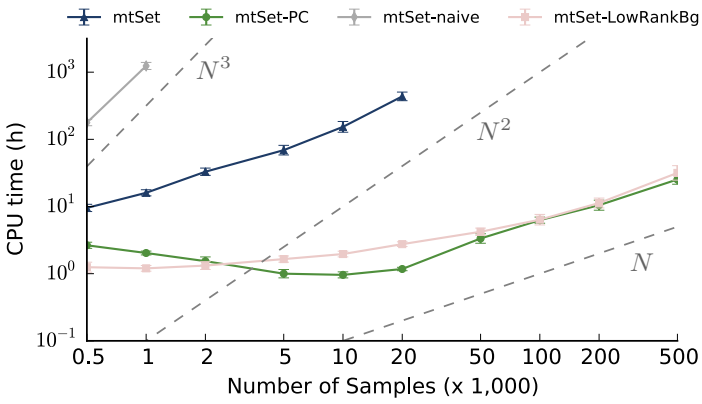
The results of the analyses are showed in **Fig. 3.1**. While a naive implementation of the algorithm becomes impractical already for analysis of cohorts with 1,000 samples, the full mtSet model can be used for genetic analyses in structured populations of moderate size (20,000 individuals). For analyses of unrelated individuals, mtSet-PC and mtSet-LowRankBg can be applied to even larger cohorts, enabling multi-trait set test analyses in cohorts with up to 500,000 individuals. Note that the decrease in CPU time for the mtSet-PC method in the range of 500-10,000 individuals is due to the decreasing number of the iterations needed by the optimisation algorithm (BFGS) to achieve convergence. This is the consequence of the fact that the log marginal likelihood is more informative (peaked) when considering bigger sample sizes.

We also assessed scalability of the considered methods when considering larger numbers of traits. Runtimes for a cohort with 1,000 individuals when varying the number of traits are reported in **Fig. B.4**.

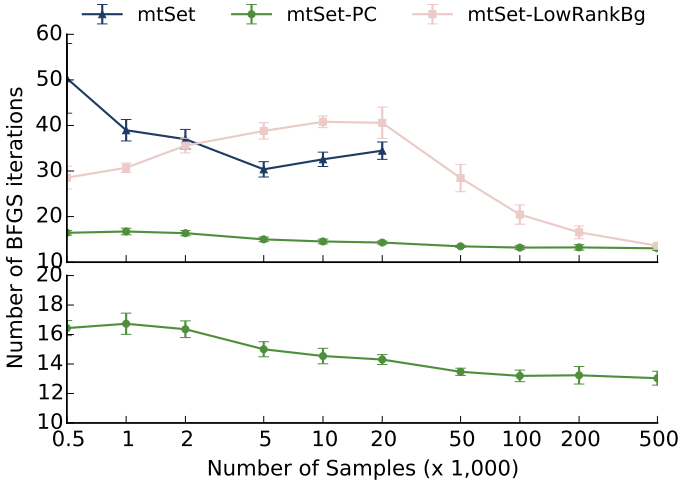
3.2.4 Calibration of P values

To assess calibration of type I error, we tested that the P values estimated by mtSet were uniformly distributed under the null hypothesis when considering cohorts with different types of confounding. To do so, we considered synthetic genotype data for independent individuals, population structure with European subpopulations and family relatedness (Section 3.2.1 and **Fig. B.3**). For each of the four datasets (the original data and the three synthetic cohorts), we considered five independent simulations. Phenotype values for four traits were generated from an empirical null model, without contribution from the set component. In each experiment, we considered a genome-wide sliding-window scan with a window size of 30kb and a step size of 15kb, which resulted in 178,695 set tests. P values were obtained from LLRs using a mixture of two chi-squared distributions (see Section 3.1.2), using five genome-wide permutations to estimated the parameters. For comparison, we also estimated P-values from the LLRs from an additional genome-wide permutation, which was not used to estimate the parameters of the test statistic distribution. Moreover, in addition to mtSet and mtSet-PC we also considered a single-trait set test (stSet) and single-variant LMMs, either considering individual traits (stLMM-SV) or performing joint tests across traits (mtLMM-SV). For mtSet-PC, we used the first 30 principal components of the RRM as fixed effects. stLMM-SV corresponds to an implementation of the fastLMM algorithm (Lippert et al., 2011) in LIMIX (Lippert et al., 2014c). mtLMM-SV is based on the equivalent version of the MTMM model (Korte et al., 2012), again implemented in LIMIX. Single-trait set tests are a special case of mtSet, which is equivalent to fastLMM-set model in Lippert et al. (2014a). Results for all scenarios and method are reported in **Table B.6** while QQ plots are shown in **Fig. B.5**.

Both mtSet and mtSet-PC were sufficiently calibrated when considering genotype data of unrelated 1000 Genomes Project individuals (**Fig. 3.2a**) or simulated cohorts either with no structure or with population structure (**Fig. B.5**). In contrast, the relatedness component of mtSet was important to ensure statistical calibration in the cohort with family relatedness (**Fig. 3.2b**, **Fig. B.5**).



(a) CPU time

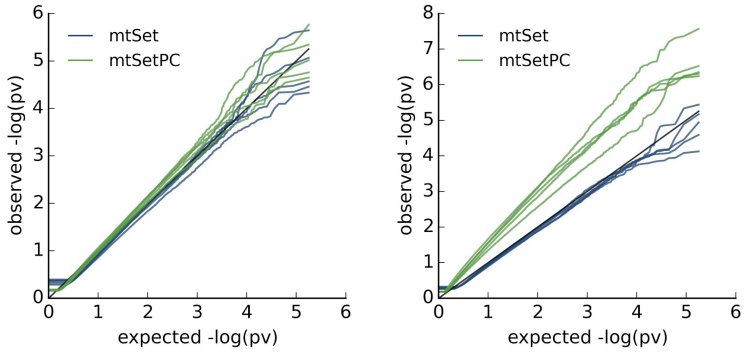


(b) Number of iterations

Figure 3.1: **Computational runtime of mtSet and alternative methods as a function of the cohort size.** Panel (**a**) shows the CPU time (h) to test associations on chromosome 20 (3,975 windows/tests) on a simulated cohort with increasing number of individuals and for four traits. Compared are the full mtSet, the PC approximation (mtSet-PC) and the low-rank approximation (mtSet-LowRank). Naive denotes the runtime for a standard LMM package, which scales cubically with the number of traits and samples. Runtime estimates were obtained on a single core of an Intel Xeon CPU E5-2670 2.60 GHz processor. Panel (**b**) shows the number of iterations of the BFGS algorithm for the different methods and cohort sizes.

3.2.5 Power comparison

To assess statistical power, we considered real genotypes from the 1000 Genomes Project, for which both mtSet and mtSet-PC yielded calibrated P values (see **Fig. 3.2a**). We considered the same set of models used in the calibration experiments. To simulate a representative range of genetic scenarios, we varied the variance explained by the causal region, the number of signal variants in the causal region, the percentage of shared variants across traits, the proportion of variance explained by genetic background, the percentage of residual variance explained by unobserved covariates, the percentage of background signal that is shared across traits and the window size (see Section 3.2.2, **Table B.4**). For each parameter setting, we simulated a total of 1,000 datasets with four phenotypes each, resulting in 44,000 experiments in total. In order to reduce the computational burden, we restricted each simulated dataset to randomised regions drawn from chromosome 20, each containing 100 windows (corresponding to genomic regions of approximately 2Mb). For set tests, we considered a sliding-window approach (30kb window size, 15kb step size). To compare alternative methods, we counted the top-associated window as a true positive if it overlapped with the simulated causal region (+/- 50kb) and was significantly associated at a family-wise error rate (FWER) of 10%. For single-variant methods the top-associated window was



(a) 1000 Genomes Project **(b)** Related individuals

Figure 3.2: **QQ-plots for the P-values obtained by mtSet and mtSet-PC considering datasets with different sample structure.** (**a**) QQ-plots for the P-values obtained by mtSet and mtSet-PC when simulating four phenotypes with only background effects (no set term) using genotypes from 1000 Genotype Project individuals. Shown are results from 5 (independent) simulations. (**b**) Identical analysis as in (**a**), considering simulated phenotype data based on synthetic genotypes with related individuals.

defined as the window centred on the lead variant with the same size as the testing window used for set tests. Single-trait methods were applied to individual traits in turn, scoring each window/variant based on the maximum test statistics across traits. The family-wise error rate was controlled for using a two-step permutation approach. For each simulated phenotype, we first repeated the experiment with permuted genotypes and recorded the maximal test statistic. In a second step, we pooled the maximal test statistics over datasets with identical simulation parameter settings. We used this distribution to define an empirical FWER as the fraction of maximal test statistics under permutations that is larger than the observed one (Westfall et al., 1993). This approach accounts for LD and the effective number of tests performed, thereby enabling an objective comparison between single-variant method and set tests.

Varying the number of causal loci. Initially, we assessed the sensitivity of different models with respect to the number of causal loci. When increasing the numbers of causal loci within the simulated region, the performance of single-locus models (stLMM-SV, mtLMM-SV) deteriorated markedly (**Fig. 3.3a**), since each locus on its own could only explain parts of the phenotypic variance explained by the region. As expected set tests were able to aggregate over multiple effects and were generally well powered when two or more causal loci were simulated.

Varying the proportion of shared background signal. As expected, when we altered the proportion of shared background effects (i.e., the parameter α introduced in Section 3.2.2), single-trait methods were not able to leverage correlated background (**Fig. 3.3b**). In contrast, multi-trait models account for trait-trait correlations and hence were substantially more powerful than the corresponding single-trait alternatives when the traits were correlated (**Fig. 3.3b**).

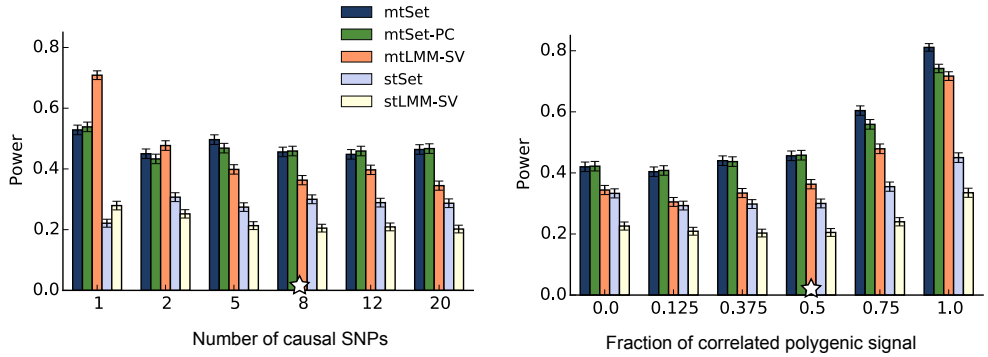
Varying the size of the testing region. We also studied the power and computational efficiency of set tests when considering different testing-region sizes (**Fig. 3.4**, **Fig. B.6**). Varying the size of the testing region from 1 kb to 200 kb, the power results were generally robust (**Fig. 3.4**). However, as expected mtSet was best powered when the size of the region term matched the simulated size of the true causal regions (30 kb). **Fig. 3.4** and **Fig. B.6** suggest that when selecting the size of the testing window both the local linkage disequilibrium and number of SNPs within regions should be considered. Testing regions that are too small will lead to high LD among the variants in the set component, which results in limited advantages of set tests compared to

single-variant LMMs. Conversely, large regions may lead to reduced power, as a consequence of including many non-causal variants in the set component, and increased computational burden, as the computational efficiency of mtSet depends on the number of variants in the set.

Varying other simulation parameters. **Fig. B.7** shows results across all simulated scenarios. We note that the exact mtSet model was in general slightly better powered than the PC-based approximation (mtSet-PC), confirming the expected benefits of including a relatedness component (Kang et al., 2010). In sum, these results show that mtSet is a robust method for association mapping across a wide range of simulated genetic architectures.

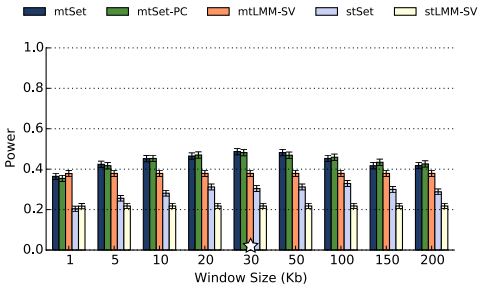
3.3 Applications to real data

To illustrate the advantages of mtSet we considered two real data application: a human GWAS of four lipid traits from the Northern Finland Birth cohort (Sabatti et al.,

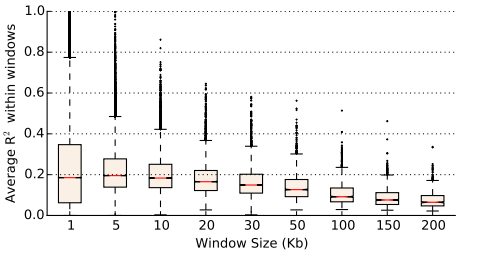


(a) **(b)**

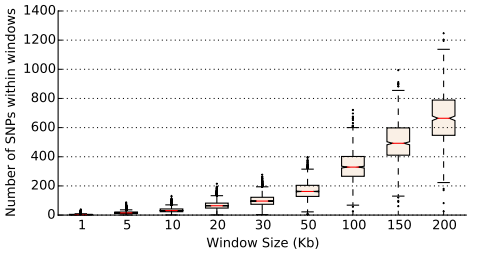
Figure 3.3: **Power comparison of alternative methods on simulated data from 1000 Genomes Project genotypes.** Power comparison of alternative methods on simulated data based on 1000 Genome Project data with four phenotypes, varying the complexity of true causal effects (number of causal variants) (**a**) and the extent of correlated genetic background (**b**). The stars denote the default values that are kept constant when varying the respective other parameter. Shown is the average power for different methods and simulation settings across 1,000 repeated experiments where the error bars denote standard errors. Compared are the full-rank multi-trait set test (mtSet), the PC approximation (mtSet-PC), the single-trait set test (stSet), a multi-trait single-variant LMM (mtLMM-SV) and a single-trait single-variant LMM (stLMM-SV).



(a) Power



(b) Average R^2 within windows



(c) Number of SNPs

Figure 3.4: **Power comparison when varying the size of the set component on simulated data from 1000 genomes project genotypes. (a)** Shown is power at 10% family-wise error rate for mtSet, stSet, mtSet-PC, mtLMM-SV and stLMM-SV when varying the window size for set test approaches. While the methods are overall robust, the set test methods were most powerful when the testing size matched the size of the simulated causal region. **(b)** Shown is the averaged squared correlation coefficient within a window as a function of the window size. Larger windows tend to contain fewer tightly correlated SNPs within windows. **(c)** Shown is the number of SNPs within windows as a function of the window size. When selecting the size of the testing window both linkage disequilibrium and number of SNPs within regions should be taken into account.

2009) and a GWAS of six haematology traits in a population of outbred rats (Baud et al., 2014). While the first is a medium-sized cohort of unrelated individuals (5,256

individuals), the second is a smaller cohort (1,334 individuals) characterised by strong genetic relatedness. The RRM of both datasets is shown in **Fig. 2.3**. **Fig. B.8** shows the distribution of the number of variants and the squared average correlation coefficient between variants in genetic regions, both as a function of the considered window size. As these figures show, the rat genotypes have a slower LD decay, which is consequence of the large haplotype blocks that characterise this dataset.

3.3.1 Genetic analysis of lipid traits in human

We applied mtSet to data from four lipid-related traits (C-reactive protein (CRP), triglycerides (TRIGL) and LDL and HDL cholesterol levels) measured in 5,256 unrelated individuals from the Northern Finland Birth cohort (Sabatti et al., 2009), which were previously considered for multi-trait analysis using single-variant LMMs (Korte et al., 2012; Zhou and Stephens, 2014). Following the approach taken in Zhou and Stephens (2014), we regressed out sex, usage of contraceptive pill (Pills31) and oral contraception (ZP4202U) and then quantile-normalised each trait to follow a unit variance normal distribution (see also Fusi et al. (2014a) for a comparison of alternative pre-processing methods in the context of LMMs). For set tests we employed a sliding-window approach, considering a window size of 100 kb and a step size of 50 kb (for a total of 328,517 variants with an MAF of at least 1%). Regions with fewer than 4 SNPs were discarded (1,802 SNPs, corresponding to <0.5%), resulting in a total of 51,658 sets for analysis. Heritability estimates and correlation coefficients on the null models of mtSet were in line with those previously reported in Korte et al. (2012) (**Table B.7** - the procedure for calculating standard errors is described in Section D.3), where small deviations are likely due to small differences in the phenotype normalisation. Following the strategy described in Section 3.1.2, we estimated P-values using 10 genome-wide permutations per window to fit an empirical null distribution. A genome-wide run required 49h for mtSet (null model: 2.58 min, average window: 5.18s), and 5h for mtSet-PC (null model: 44.89s, average window: 1.78 s). We again compared mtSet and mtSet-PC to a single-trait set test, and single-trait and multi-trait LMMs for single-variant testing. All methods yielded well-calibrated P values (**Fig. B.9-B.10**). Significance of QTLs was assessed at the Bonferroni adjusted significance level $\alpha < 0.01$. For single-trait methods, we considered the minimum P-values across traits, again adjusting for the additional tests using Bonferroni.

Manhattan plots for all methods are shown in **Fig. B.11** and a tabular summary of the QTLs is provided in **Table B.8**. Notably, mtSet identified 14 genome-wide significant quantitative trait loci (QTL) (alpha<0.01, Bonferroni adjusted), 13 of which

have previously been identified in a larger meta-analysis (Teslovich et al., 2010) and the remaining one has been reported when applying single-variant multi-trait LMMs to the same dataset (Zhou and Stephens, 2014). Single-variant LMMs missed four associations and single-trait set tests failed to detect three of the associations detected by mtSet. In contrast, mtSet identified all but one association found when considering the union of associations detected by previous methods and retrieved one additional association close to the *ANGPTL3* (Anglopoletin-like 3) gene, a known regulator of lipid metabolism in mice (Koishi et al., 2002). Notably, mtSet-PC was even slightly better powered than mtSet (identifying 16 QTLs). The model retrieved all associations found by mtSet or any other method and found an additional association close to the *LCAT* (lecithin cholesterol acyltransferase) gene, which is known to contribute with multiple rare alleles to low plasma levels (Cohen et al., 2004). Finally, in order to assess robustness of the results to the choice of the window size we repeated the analysis considering either a smaller window size of 60kb or a larger window size of 300kb. **Table B.9** shows that results are overall robust when considering different window sizes.

3.3.2 Genetic analysis of haematology traits in rat

To evaluate mtSet in a setting with strong relatedness, we considered a QTL study of 1,334 outbred rats (The Rat Genome Sequencing and Mapping Consortium, 2013) and applied mtSet to six traits related to basal haematology (concentrations of basophils (basos), eosinophils (eos), large unstained cells (luc), lymphocytes (lymphs), monocytes (monos) and neutrophils (neuts)). We regressed out sex and batch covariates and quantile normalised each trait to a unit variance normal distribution. Variants were filtered to have a minor allele frequency (MAF) of at least 5% resulting in 4,138,000 variants for the analysis. Because of the large haplotype blocks in this particular study (multi-parent cross genetic design), we considered larger regions (1Mb size) and a step size of 500kb, resulting in a total of 5,220 sets. Heritability estimates from a single-trait LMM were consistent with the marginal heritability estimates of the mtSet null model (**Table B.10**). To estimate P-values, we used 30 genome-wide permutations per window to fit an empirical null distribution.

First, to study calibration of P values using different correction strategies we compared mtSet to mtSet-PC and a single-variance component model without correction for population structure (mtSet-NoBg). All three models were calibrated when permuting the SNPs within the set (using genome-wide permutations to retain the LD structure, see Section 3.1.2), which corresponds to empirical data from the null without asso-

ciation signal (**Fig. 3.5a**). However, for the observed data (**Fig. 3.5b**, **Fig. B.13**), only mtSet yielded calibrated results, confirming the expected benefits of the second variance component term to control for relatedness (Kang et al., 2010; Schifano et al., 2012). The QQ-plots in **Fig. 3.5c** were obtained after removing duplicate SNPs, i.e. only considering unique variants.

We then compared mtSet to different LMMs, including single-trait set test, single-trait LMM for single variants and multi-trait-LMM for single variants. Again, significance was assessed at FWER < 0.01 significance level, adjusting for multiple testing using Bonferroni (considering only unique variants in the dataset). For single-trait methods we corrected Bonferroni both across variants and traits. **Fig. 3.5c** shows the Manhattan plots for the four methods (for single-trait methods we report the minimum P value across traits; Manhattan plots for single-traits tests are reported in **Fig. B.13**). A tabular summary of the results is given in **Table B.11**. mtSet identified one additional QTL (alpha < 0.01, Bonferroni adjusted, see **Fig. 3.5c**). This QTL points to NFKB2, a gene that is involved in immune response in humans (Wit et al., 1998), making it also a plausible candidate gene for haematological traits in rat.

3.4 Summary and discussion

In this chapter, I have shown how set tests can be extended to enable joint analysis of multiple traits. Through applications to both simulated and real data, I have shown that the mtSet model increases power compared to existing LMMs for association testing, while retaining computationally efficiency and controlling for arbitrary confounding structure. mtSet builds on an efficient inference scheme for multi-trait LMMs with two variance components, by exploiting the low rank structure of genetic region terms. The proposed algebraic implementation extends the efficient inference schemes of existing LMMs for genome-wide association studies.

Although mtSet can be used to fit variance component models in general, it is not free of limitations. First, it builds on the assumption that the number of variants in the analysed region is lower than the number of samples. Although this assumption holds for many use cases, analysis of high-density genotype markers or genome sequencing data will entail a trade-off between the size of the genetic region that can be tested and computational efficiency. Notably, as the marker density increases, one will typically test smaller regions to improve the resolution of the mapping, which may help mitigate this concern.

Additionally, the efficient inference scheme of the mtSet model requires fully ob-

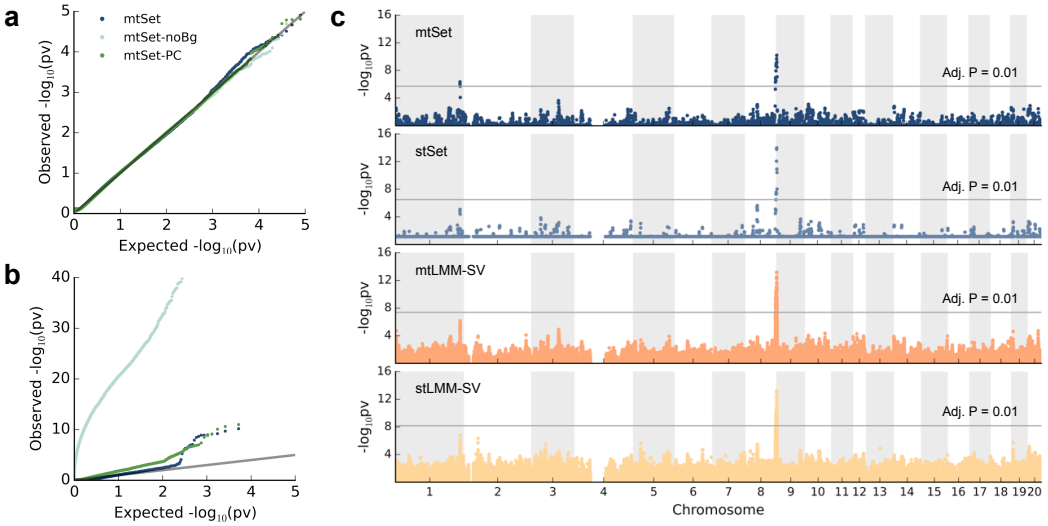


Figure 3.5: **Application of mtSet to six basal haematology traits in the rat data.** (**a**,**b**) Comparison of the calibration of alternative methods, considering both permuted (null) data (**a**) and the real rat dataset (**b**). The null data is obtained from a single permutation applied to SNPs used to estimate the set component. Compared are mtSet run with relatedness component (mtSet), without relatedness component, principal components to adjust for structure (mtSet-PC) and a variant of mtSet without relatedness component and no PCs (mtSet-NoBg). (**c**) From top to bottom, Manhattan plots for mtSet, the single-trait set test (stSet), a multi-trait single-variant LMM (mtLMM-SV) and a standard single-trait single-variant LMM (stLMM-SV). For single-trait models (stLMM-SV, stSet), minimal P-values across traits are shown. The horizontal line indicates the FWER = 0.01 significance threshold (Bonferroni adjusted).

served phenotype data. This assumption is shared with the majority of alternative inference schemes to efficiently fit LMMs (Zhou and Stephens, 2014; Furlotte and Eskin, 2015). If phenotype data are partially observed, phenotype imputation methods, e.g. the one proposed by Dahl et al. (2016), can be reused.

Although I here exclusively considered tests for shared effects across multiple traits (i.e., we used a rank-one parametrisation of the set trait covariance), the mtSet model can describe more complex genetic architectures as I discuss in the next chapter.

4 | Testing for polygenic interactions using set tests

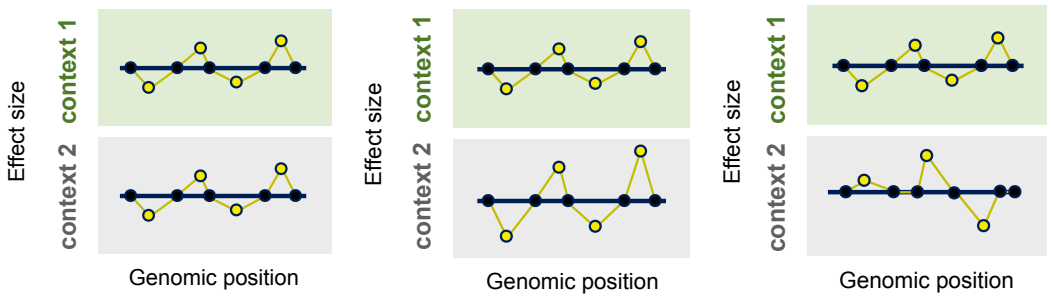
Genotype-context (GxC) interactions are genetic effects that depend on external contexts, including environment. Studies that explicitly test for differences in genetic effects have revealed that GxC interactions are common and occur for a large number of complex traits, including whole-body phenotypes, disease susceptibilities and psychiatric disorders (Warren et al., 2012; Modinos et al., 2013; Winkler et al., 2015; Young et al., 2016). While for global phenotypes typical contexts under study are environmental factors, genetic effects on molecular traits have been surveyed across multiple tissues, cell types and stimuli (see also Section 1.2.4).

There are two canonical designs for the analysis of genotype-context interactions. If the system permits the study of the same genotype in different conditions, the same individuals can be phenotyped repeatedly in different contexts. However, in general this design is not always feasible, for example, when studying variation of global phenotypes in human. In this case a typical approach is to stratify the population into distinct subgroups using a context variable. In the following, I will refer to these two alternative study designs as *complete* and *stratified* designs.

Building on the mtSet model from the previous chapter, I here derive interaction tests between sets of genetic variants and categorical contexts (iSet). This approach generalises previous single-variant and set-based interaction tests and can be applied to analyse datasets with either complete or stratified designs. iSet offers two major advantages compared to the classical single-variant GxC test that uses a fixed effect to test for differential effect sizes between contexts. First, by accounting for effects due to multiple causal variants the method offers increased statistical power for detecting polygenic interactions, analogous to set tests for persistent genetic effects (Listgarten et al., 2013; Casale et al., 2015; Brown et al., 2016). Second, as described in more detail in the following, this approach enables the characterisation of changes in the genetic

architecture at specific loci across the analysed contexts.

Let us consider a categorical context and a quantitative trait locus harbouring multiple causal variants. The baseline setting is that the effects of the causal variants at the locus do not depend on the analysed context (**Fig. 4.1a**). In other words there is no GxC; we denote such genetic effects *persistent effects*. A simple model of interaction at the locus assumes that the local genetic effects in one context are proportional to the effects in all other contexts; a criterion that has also been considered to assess co-localisation of genetic effects across multiple traits (Wallace, 2013) (**Fig. 4.1b**). We denote this class of interactions *rescaling-GxC*. In the most general case, however, there may also be changes in the configuration of causal variants between contexts (**Fig. 4.1c**). We denote this class of interactions *heterogeneity-GxC*. The proposed iSet method can distinguish between these two classes of interactions.



(a) Persistent effects **(b)** Rescaling-GxC effects **(c)** Heterogeneity-GxC effects

Figure 4.1: **Schematic representation of different architectures of polygenic effects.** Alternative genetic architectures that are modelled in iSet: (**a**) persistent effects, where causal variants have identical effects in the two contexts, (**b**) rescaling-GxC effects, where the effects of causal variants in the two contexts are proportional, (**c**) and heterogeneity-GxC effects, where the effect sizes in the two contexts are not simply proportional, indicating a change in the configuration of causal variants.

In Section 4.1, I introduce the iSet model. In Sections 4.2-4.3, I describe results from simulation experiments and an application to a stimulus eQTL study. In Section 4.4, I present the extension of iSet to enable analyses of stratified designs and I illustrate this approach in both simulations and a genotype-sex interaction analysis from real data. Finally, in Section 4.5, I present a summary and a discussion of the limitations and possible extensions of iSet.

4.1 The interaction set test

In this section, I derive the different models considered by iSet from a generative linear model perspective. I then discuss the proposed statistical tests, the strategy to obtain P values and an interpretation of the estimated variance parameters. Initially, we will study the case of complete designs. The generalisation of the model to stratified designs will be given in Section 4.4.

4.1.1 Model derivation

The $N \times C$ phenotype matrix $\boldsymbol{Y}$ for N individuals and two or more contexts C is modelled as the sum of the contribution from K fixed effect covariates, the contribution from R variants in the region of interest (set component), a term accounting for population structure or relatedness (relatedness component) and residual noise

$$\boldsymbol{Y} = \underbrace{\boldsymbol{F}\boldsymbol{B}}_{\text{fixed effects}} + \underbrace{\boldsymbol{G}\boldsymbol{W}}_{\text{set component}} + \underbrace{\boldsymbol{U}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{\Psi}}_{\text{noise}}. \tag{4.1}$$

Here, $\boldsymbol{G} \in \mathbb{R}^{N \times R}$ denotes the standardised genotype matrix of the R genetic variants in the set, $\boldsymbol{W} \in \mathbb{R}^{R \times C}$ the matrix of their effect sizes across the C contexts, $\boldsymbol{F} \in \mathbb{R}^{N \times K}$ the design matrix of the K covariates, $\boldsymbol{B} \in \mathbb{R}^{K \times C}$ the matrix of their effect sizes, and $\boldsymbol{U}_g$ and $\boldsymbol{\Psi}$ are random effects that account for relatedness and noise and follow matrix-variate normal distributions:

$$\boldsymbol{U}_g \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{C}_g, \boldsymbol{R}\right) \quad \text{and} \quad \boldsymbol{\Psi} \sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{C}_n, \boldsymbol{I}_N\right), \tag{4.2}$$

where $\boldsymbol{R} \in \mathbb{R}^{N \times N}$ denotes the RRM and $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ are $C \times C$ covariance matrices accounting for co-variation of trait measurements across contexts due respectively to the contributions from relatedness and noise.

Let us consider a normal prior on the effect sizes of each variant in the region

$$\boldsymbol{W}_{s,:} \sim \mathcal{N}\left(\boldsymbol{0}, \frac{1}{R}\boldsymbol{C}_r\right), \quad \forall s \in \{1, \ldots, S\}, \tag{4.3}$$

where $\boldsymbol{C}_r$ is the $C \times C$ covariance matrix of the effect sizes of the set variants across

traits. Marginalizing out $\boldsymbol{W}$, we obtain the mtSet model in Eq. (3.3)

$$\operatorname{vec}(\boldsymbol{Y}) \sim \mathcal{N}\left(\underbrace{\operatorname{vec}(\boldsymbol{FB})}_{\text{fixed effect covariates}} , \underbrace{\boldsymbol{C}_r \otimes \boldsymbol{R}_r}_{\text{set component}} + \underbrace{\boldsymbol{C}_g \otimes \boldsymbol{R}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{C}_n \otimes \boldsymbol{I}_N}_{\text{noise}} \right) \quad (4.4)$$

where $\boldsymbol{R}_r = \frac{1}{R}\boldsymbol{G}\boldsymbol{G}^\top \in \mathbb{R}^{N\times N}$ is the a local RRM, estimated solely from the R variants in the region.

A key insight derived here is that different local architectures (**Fig. 4.1**) correspond to alternative assumptions on the structure of the trait-context covariance $\boldsymbol{C}_r$ (see also **Fig. 4.2a-b**). To simplify the notation, we focus on the case of two contexts ($C = 2$). The generalisation to the analysis of multiple contexts is discussed at the end of this section.

Persistent effect model. In the scenario of local persistent effects, the matrix of the variant effect sizes can be written as

$$\boldsymbol{W} = [a\boldsymbol{\gamma}, a\boldsymbol{\gamma}] = \boldsymbol{\gamma} a \mathbf{1}_2^\top, \quad (4.5)$$

where the effect size profile $\boldsymbol{\gamma} \in \mathbb{R}^R$ of the genetic variants has same scale a in the both contexts (**Fig. 4.1a**). Considering the prior $\boldsymbol{\gamma} \sim \mathcal{N}(0, \frac{1}{R}\boldsymbol{I}_R)$ and marginalising out $\boldsymbol{\gamma}$ out, we obtain the model in Eq (4.4) with $\boldsymbol{C}_r = a^2\mathbf{1}_{2\times 2}$ (block covariance). One way to derive this result is to consider $\boldsymbol{r} = \operatorname{vec}(\boldsymbol{GW}) = \operatorname{vec}\left(\boldsymbol{G}\boldsymbol{\gamma} a \mathbf{1}_2^\top\right) = a\,(\mathbf{1}_2 \otimes \boldsymbol{G})\,\boldsymbol{\gamma}$. As $\boldsymbol{\gamma} \sim \mathcal{N}(0, \frac{1}{R}\boldsymbol{I}_R)$ it follows

$$\boldsymbol{r} \sim \mathcal{N}\left(\mathbf{0}, \underbrace{a^2\mathbf{1}_2\mathbf{1}_2^\top}_{\boldsymbol{C}_r} \otimes \underbrace{\frac{1}{R}\boldsymbol{G}\boldsymbol{G}^\top}_{\boldsymbol{R}_r} \right). \quad (4.6)$$

Rescaling-GxC model. In the scenario of rescaling-GxC effects, we consider

$$\boldsymbol{W} = [a_1\boldsymbol{\gamma}, a_2\boldsymbol{\gamma}] = \boldsymbol{\gamma}\boldsymbol{a}^\top, \quad (4.7)$$

where the effect size profile $\boldsymbol{\gamma} \in \mathbb{R}^R$ of the genetic variants has scales $\boldsymbol{a} = [a_1, a_2]$ in the two contexts (**Fig. 4.1b**). Considering the prior $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \frac{1}{R}\boldsymbol{I}_R)$ and marginalising $\boldsymbol{\gamma}$ out we obtain $\boldsymbol{C}_r = \boldsymbol{a}\boldsymbol{a}^\top$ (rank-one covariance).

This model can capture three different biological settings:

- $a_1 a_2 > 0$, the signal from the set has same direction;

- $a_1 a_2 < 0$, the signal from the set has opposite direction;
- $a_1 \approx 0$ and $a_2 \neq 0$ (or vice versa), the signal from the set is specific to one of the two contexts.

Moreover, note that for $a_1 = a_2$ the model reduces to the case of persistent effect.

General-GxC model. For the most general case in **Fig. 4.1c**, we introduce effect size profiles $\boldsymbol{\gamma}^{(1)} \in \mathbb{R}^R$ and $\boldsymbol{\gamma}^{(2)} \in \mathbb{R}^R$ and set

$$\boldsymbol{W} = \left[a_{11}\boldsymbol{\gamma}^{(1)} + a_{12}\boldsymbol{\gamma}^{(2)},\ a_{21}\boldsymbol{\gamma}^{(1)} + a_{22}\boldsymbol{\gamma}^{(2)}\right] = \left[\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}\right] \boldsymbol{A}^\top, \tag{4.8}$$

where $\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ is a rescaling matrix (a_{ij} is the scale of variant effect profile $\boldsymbol{\gamma}_j$ in context i). Introducing the prior $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \sim \mathcal{N}(0, \frac{1}{R}\boldsymbol{I}_R)$ and marginalising $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ out, we obtain the model in Eq (4.4) with $\boldsymbol{C}_r = \boldsymbol{A}\boldsymbol{A}^\top$ (full-rank covariance). Note that the model includes the cases of both rescaling-GxC and persistent effects as special cases.

Analyses across more than two contexts. For analyses of $C > 2$ contexts we can consider a model with $L \leq C$ distinct genetic signals $\left\{\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(\mathrm{L})}\right\}$. Introducing $\boldsymbol{\Gamma}_L = \left[\boldsymbol{\gamma}^{(1)}, \ldots, \boldsymbol{\gamma}^{(\mathrm{L})}\right] \in \mathbb{R}^{R \times L}$ and the rescaling matrix $\boldsymbol{A}_L \in \mathbb{R}^{L \times C}$, we can write

$$\boldsymbol{W} = \boldsymbol{\Gamma}_L \boldsymbol{A}_L. \tag{4.9}$$

Setting $\boldsymbol{\Gamma}_L \overset{\text{iid}}{\sim} \prod \mathcal{N}(0, 1)$ we obtain $\boldsymbol{C}_r = \boldsymbol{A}_L \boldsymbol{A}_L^\top$, which has rank L. This result also clarifies the aforementioned interpretation of the rank of the genetic trait-correlation matrix $\boldsymbol{C}_r$ (Section 3.1.1). The rank of $\boldsymbol{C}_r$ is the number of distinct local genetic signals across the different contexts. Note that each signal may be polygenic, resulting from the joint effect of multiple variants in the set.

4.1.2 Statistical testing

Model comparisons of the LMM in Eq (4.4) considering alternative covariance structures for the set trait covariance enables us to test for different hypothesis on the local genetic architecture (**Fig. 4.2**). Specifically, we consider the following tests:

- **Association test (mtSet).** The full-rank covariance model is tested against a

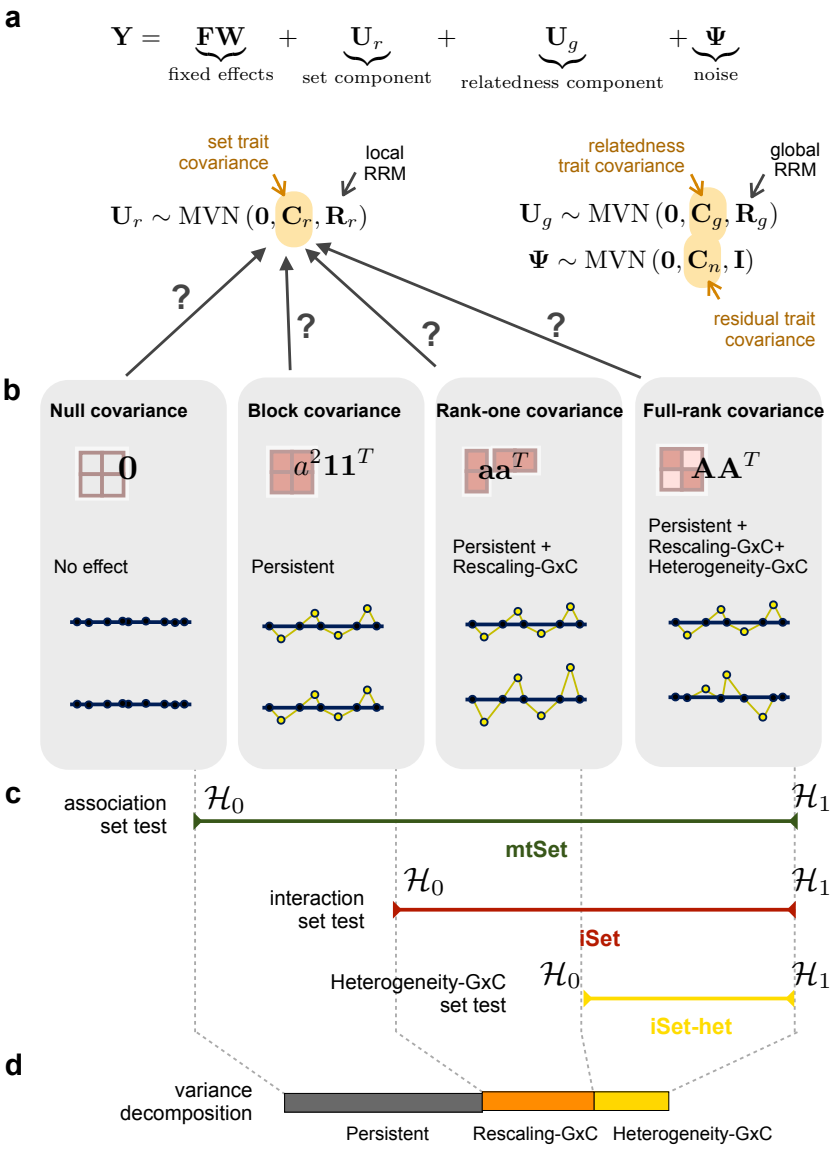


Figure 4.2: **Illustration of the iSet model.** (**a**) shows the multivariate mixed-model underlying iSet. (**b**) illustrates the different covariance models of the set trait covariance $\boldsymbol{C}_r$ considered in iSet and the corresponding changes in the local genetic architecture associated to these models. (**c**) shows the null and the alternative hypotheses that define the tests for association (mtSet), interaction (iSet) and heterogeneity-GxC (iSet-het). Interpreting the variance parameters estimated by iSet, we can quantify the proportion of local variance explainable by persistent, rescaling-GxC and heterogeneity-GxC effects (**d**).

null covariance model (no association):

$$\mathcal{H}_1 : \boldsymbol{C}_r = \boldsymbol{A}\boldsymbol{A}^\top \quad \text{vs} \quad \mathcal{H}_0 : \boldsymbol{C}_r = \boldsymbol{0} \tag{4.10}$$

- **Interaction test (iSet).** The full-rank covariance model is tested against a block covariance model (which only captures persistent effects):

$$\mathcal{H}_1 : \boldsymbol{C}_r = \boldsymbol{A}\boldsymbol{A}^\top \quad \text{vs} \quad \mathcal{H}_0 : \boldsymbol{C}_r = a^2\boldsymbol{1}_{2\times 2} \tag{4.11}$$

- **Heterogeneity-GxC test (iSet-het).** The full-rank covariance model is tested against a rank-one covariance model (which captures both persistent and rescaling-GxC effects):

$$\mathcal{H}_1 : \boldsymbol{C}_r = \boldsymbol{A}\boldsymbol{A}^\top \quad \text{vs} \quad \mathcal{H}_0 : \boldsymbol{C}_r = \boldsymbol{a}\boldsymbol{a}^\top \tag{4.12}$$

Note that the covariance models introduced in the previous section are nested and thus statistical tests are well defined. A summary of the different covariance structures considered by the model and their use to derive specific tests is shown in **Fig. 4.2.**

P values. For the proposed tests, we consider the log-likelihood ratio (LLR) test statistics. While for the standard association test (mtSet) P values can be obtained using the permutation-based approach described in Section 4.1.2, permutation schemes are not defined for interaction tests (Bůžková et al., 2011). Following Bůžková et al. (2011), we consider a parametric bootstrap procedure to estimate P values for the iSet and the iSet-het tests. To generate test statistics from an empirical null, this procedure consists in drawing phenotypes from the null model with parameter values that maximise the likelihood on real data. Similar to the strategy employed for mtSet, we consider a small number of parametric bootstraps for each region (typically 10-100 bootstraps) and pool the obtained null LLRs across all tested regions. We then use the estimated distribution of null LLRs to obtain empirical P values.

In an analysis of T genomic regions, the procedure to obtain P values for the the three tests can be summarised as follows:

- For each of the T sets
  - fit the no-association model ($\mathcal{H}_{\text{na}}$), the block covariance model ($\mathcal{H}_{\text{block}}$), the rank-one covariance model ($\mathcal{H}_{\text{rank1}}$) and the full-rank covariance model ($\mathcal{H}_{\text{full}}$) and estimate LLRs for mtSet ($\mathcal{H}_{\text{full}}$ vs $\mathcal{H}_{\text{na}}$), iSet ($\mathcal{H}_{\text{full}}$ vs $\mathcal{H}_{\text{block}}$) and iSet-het ($\mathcal{H}_{\text{full}}$ vs $\mathcal{H}_{\text{rank1}}$);

– sample J LLRs from the null hypothesis for each of the three tests

  * for mtSet, null LLRs are sampled as the LLRs from the mtSet test considering J permutations the individuals in the set component;
  * for iSet, null LLRs are sampled as the LLRs from the iSet test considering J parametric bootstraps from $\mathcal{H}_{\text{block}}$;
  * for iSet-het, null LLRs are sampled as the LLRs from the iSet-het test considering J parametric bootstraps from $\mathcal{H}_{\text{rank1}}$;

- for each of the three tests, pool the JT null LLRs across regions to obtain an empirical null. Empirical P values are obtained as the fraction of null LLRs that are at least as extreme as the observed one.

4.1.3 Interpretation of the variance parameters

The expected sample variance of a random vector $\boldsymbol{u}$ following a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{K}$, $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, is (Searle, 1982, p 67)

$$\mathbb{E}\left[\text{var}\left(\boldsymbol{u}\right)\right] = \frac{\text{tr}\left(\boldsymbol{P}_n \boldsymbol{K}\right)}{n-1}, \tag{4.13}$$

where n is the total number of samples and $\boldsymbol{P}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_{n\times n}$ is the centring matrix.

Using Eq (4.13), iSet allows for estimating (i) the fraction of variance explained by the genetic region (as in Kostem and Eskin, 2013) and (ii) the relative proportions of local variance that is explainable by persistent, rescaling-GxC and heterogeneity-GxC effects (**Fig. 4.2d**).

Fraction of variance explained by the region. Considering the model in Eq (4.4), the variance explained by the set component (v_{set}), the relatedness (v_{rel}) and the noise component (v_{noise}) across all observations are

$$v_{\text{set}} = \frac{1}{NC-1}\text{tr}\left(\boldsymbol{P}_{NC}\left(\hat{\boldsymbol{C}}_r \otimes \boldsymbol{R}_r\right)\right) \tag{4.14}$$

$$v_{\text{rel}} = \frac{1}{NC-1}\text{tr}\left(\boldsymbol{P}_{NC}\left(\hat{\boldsymbol{C}}_g \otimes \boldsymbol{R}_g\right)\right) \tag{4.15}$$

$$v_{\text{noise}} = \frac{1}{NC-1}\text{tr}\left(\boldsymbol{P}_{NC}\left(\hat{\boldsymbol{C}}_n \otimes \boldsymbol{I}_N\right)\right). \tag{4.16}$$

Here $\hat{\boldsymbol{C}}_r$, $\hat{\boldsymbol{C}}_g$ and $\hat{\boldsymbol{C}}_n$ denote the maximum-likelihood estimator (MLE) of $\boldsymbol{C}_r$, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ considering a full-rank covariance model for $\boldsymbol{C}_r$. The expected proportion of

variance explained by the genetic region can be estimated as

$$h^2_{\text{set}} = \frac{v_{\text{set}}}{\text{var}(\text{vec}\,(\boldsymbol{FB})) + v_{\text{set}} + v_{\text{rel}} + v_{\text{noise}}}. \tag{4.17}$$

Note that v_{set}, v_{rel} and v_{noise} can be computed efficiently as

$$\boldsymbol{P}_{NC} = \boldsymbol{I}_{NC} - \frac{1}{NC}\mathbf{1}_{NC\times NC} = \boldsymbol{I}_C \otimes \boldsymbol{I}_N - \frac{1}{NC}\left(\mathbf{1}_{C\times C} \otimes \mathbf{1}_{N\times N}\right). \tag{4.18}$$

Decomposing the local variance. To estimate the relative proportions of local variance that is explainable by persistent, rescaling-GxC and heterogeneity-GxC effects we use the following strategy:

- consider the block and the low-rank approximation of $\hat{\boldsymbol{C}}_r$

$$\hat{\boldsymbol{C}}_r^{(\text{block})} = \text{mean}(\hat{\boldsymbol{C}}_r)\mathbf{1}_{C\times C} \tag{4.19}$$

$$\hat{\boldsymbol{C}}_r^{(\text{lr})} = \lambda \boldsymbol{v}\boldsymbol{v}^\top, \tag{4.20}$$

  where λ is the largest eigenvalue of $\hat{\boldsymbol{C}}_r$ and $\boldsymbol{v}$ the corresponding eigenvector;

- estimate the expected sample variance of the random vectors $\boldsymbol{u}^{(\text{block})} \sim \mathcal{N}\left(\mathbf{0}\,,\, \hat{\boldsymbol{C}}_r^{(\text{block})} \otimes \boldsymbol{R}_r\right)$ and $\boldsymbol{u}^{(\text{lr})} \sim \mathcal{N}\left(\mathbf{0}\,,\, \hat{\boldsymbol{C}}_r^{(\text{lr})} \otimes \boldsymbol{R}_r\right)$, which we denote with v_{pers} and v_{lr} respectively, using Eq (4.13);

- define the variance explained by persistent, rescaling-GxC and heterogeneity-GxC as v_{pers}, $v_{\text{lr}} - v_{\text{pers}}$, $v_{\text{set}} - v_{\text{lr}}$[1] .

An alternative strategy to estimate these variances is to estimate the variance explained by persistent and rescaling-GxC effects directly from the MLE of $\boldsymbol{C}_r$ when considering block and low-rank models for $\boldsymbol{C}_r$. Empirically, we observed that two approaches gave very similar results (data not shown).

4.1.4 Relationship to existing interaction tests

iSet extends existing multivariate LMMs and set-based interaction tests (see **Table C.1** for a tabular comparison). While multivariate LMMs have been limited to analyses of single variants, existing interaction set tests build on a univariate LMM and cannot be applied to analyse datasets with complete designs. In the following, I provide a brief overview of these methods.

[1]Note that the defined variances are well defined (i.e., non-negative) as $v_{\text{pers}} \leq v_{\text{lr}} \leq v_{\text{set}}$.

Multi-trait LMMs for interaction test. In studies with complete designs, it is possible to use the multi-trait linear mixed model in Eq (2.79) to test for GxC interactions by regarding trait measurements across the different contexts as multiple phenotypes. In this setting, a specific-effect test (see Section 2.4.3) corresponds to a test for interaction. For stratified designs considering a categorical context (i.e., two subgroups), the corresponding model can be cast as

$$\boldsymbol{y} = \underbrace{\boldsymbol{F}\boldsymbol{b}}_{\text{covariates}} + \underbrace{\boldsymbol{e}\alpha}_{\text{context}} + \underbrace{\boldsymbol{g}\beta}_{\text{genetic variant}} + \underbrace{(\boldsymbol{g} \odot \boldsymbol{e})\,\gamma}_{\text{GxC}} + \underbrace{\boldsymbol{u}}_{\text{confounding}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}}, \tag{4.21}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ is the phenotype vector for N individuals, $\boldsymbol{F} \in \mathbb{R}^{N \times K}$ is the fixed effect design matrix for K covariates, $\boldsymbol{e} \in \mathbb{R}^N$ a binary indicator that specifies the context for each individual and $\boldsymbol{g} \in \mathbb{R}^N$ the genotype vector of the variant being tested. Additionally, $\boldsymbol{b} \in \mathbb{R}^K$, α, β and γ denote the fixed effects of the covariates, the context variable, the variant being tested and the GxC interaction, respectively. Finally, $\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_g^2 \boldsymbol{R}\right)$ is a random effect modelling genetic relatedness and $\boldsymbol{\psi} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_n^2 \boldsymbol{I}_N\right)$ is the residual noise, where $\boldsymbol{R} \in \mathbb{R}^{N \times N}$ denotes the realised relatedness matrix (RRM) and σ_g^2 and σ_n^2 the relatedness and noise variance components. The test $\gamma \neq 0$ allows assessing the presence of a GxC interaction.

Set-based interaction tests. One of the first set tests is the Turkey's one degree-of-freedom (dof) test (Chatterjee et al., 2006). Using the notation introduced above and denoting with $\{\boldsymbol{g}_1, \dots \boldsymbol{g}_R\}$ the genotype vectors for the R variants in the analysed region, Chatterjee et al. (2006) considered the model

$$\boldsymbol{y} = \underbrace{\boldsymbol{F}\boldsymbol{b}}_{\text{covariates}} + \underbrace{\boldsymbol{e}\alpha}_{\text{context}} + \underbrace{\sum_s \boldsymbol{g}_s \beta_s}_{\text{set}} + \underbrace{\sum_s (\boldsymbol{g}_s \odot \boldsymbol{e})\,\gamma_i}_{\text{set GxC}} + \underbrace{\boldsymbol{\psi}}_{\text{noise}}, \tag{4.22}$$

with the assumption that the interaction effect for variant i is proportional to its marginal effect (i.e., $\gamma_i = \theta\beta_i$). The presence of GxC interactions can be assessed by considering the one dof test $\theta \neq 0$. For analysis of binary phenotypes, Jiao et al. (2013) proposed an alternative strategy to re-weight the interaction effects of the variants in the set based on the correlation between the set genotypes and the context variable.

An alternative strategy to aggregate GxC effects across multiple variants is to use a random effect model. Different random-effect models for interaction set test have been proposed (Tzeng et al., 2011; Lin et al., 2013; Lin et al., 2016; Zhao et al., 2015), all of which build on closely related statistical models. In the following, I describe the

Gene-Environment Set Association Test (GESAT) (Lin et al., 2013), a representative interaction set test that we consider as a comparison partner (Section 4.4). The GESAT model is similar to the model in Eq (4.22); however, the presence of interaction is assessed by modelling $\boldsymbol{\gamma}$ as a random effect, $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \tau \boldsymbol{I}_R)$, and testing $\tau \neq 0$. This is done by using a score test, similar to Wu et al. (2011). Score tests are attractive, as they do not require to explicitly fit the alternative model. When considering sets with a large numbers of variants, the number of fixed effects in the null model is also large, which may lead to overfitting. To overcome this issue, Lin et al. (2013) have considered ridge regression to fit the null model (Hoerl and Kennard, 1970)[2]. The score test statistics is

$$\boldsymbol{Q} = (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^\top \boldsymbol{S}\boldsymbol{S}^\top (\boldsymbol{y} - \hat{\boldsymbol{\mu}}), \tag{4.23}$$

where $\boldsymbol{S} = [\boldsymbol{g}_1 \odot \boldsymbol{e}, \ldots, \boldsymbol{g}_R \odot \boldsymbol{e}]$ and $\hat{\boldsymbol{\mu}}$ is the optimised mean under the null model (i.e., for $\boldsymbol{\gamma} = 0$). It can be shown that $\boldsymbol{Q}$ follows a mixture of χ^2 distributions with 1 dof (Wu et al., 2011) and that the coefficients of this mixture are the eigenvalues of the matrix

$$\boldsymbol{T} = \boldsymbol{P}_0^{\frac{1}{2}} \boldsymbol{S}\boldsymbol{S}^\top \boldsymbol{P}_0^{\frac{1}{2}} \in \mathbb{R}^{N \times N}, \tag{4.24}$$

where $\boldsymbol{P}_0 = \hat{\sigma}_n^2 \boldsymbol{I}_N - \hat{\sigma}_n^2 \tilde{\boldsymbol{F}} \left(\tilde{\boldsymbol{F}}^\top \tilde{\boldsymbol{F}} \right)^{-1} \tilde{\boldsymbol{F}}^\top$ and $\hat{\sigma}_n^2$ is the maximum likelihood estimate of the noise variance under the null and $\tilde{\boldsymbol{F}} = [\boldsymbol{F}, \boldsymbol{e}, \boldsymbol{g}_1, \ldots, \boldsymbol{g}_R] \in \mathbb{R}^{N \times (K+R+1)}$. P values can be calculated from the score test statistics under the assumption that the asymptotic distribution is valid, typically using the Davies method (Davies, 1980). Computation of the eigenvalue decomposition of $\boldsymbol{T}$ has complexity $O(N^3)$. However, if the number of variants is lower than the number of samples ($R < N$) the non-zero eigenvalues of $\boldsymbol{T}$ can be computed as the eigenvalues of $\boldsymbol{S}^\top \boldsymbol{P}_0 \boldsymbol{S}$, whose computation and eigenvalue decomposition requires $O(NR^2)$ and $O(R^3)$ respectively. For details on score-based methods I refer to Wu et al. (2011), Lee et al. (2012c) and Lippert et al. (2014a).

Building on a multi-trait framework, iSet extends existing interaction set tests: (i) it enables interaction set test in the analysis of data with either complete or stratified designs, (ii) it explicitly model noise heterogeneity across the different contexts, (iii) it allows for testing different classes of GxC effects, (iv) it enables estimation of variance components. On the other hand, some of the methods mentioned here have features that are not available in iSet. Specifically, the model is not designed for rare variant as-

[2]Ridge regression is a penalised maximum likelihood method. Specifically, ridge regression considered a quadratic penalisation on the fixed effects of the model. The extent of the penalisation is selected by cross-validation.

sociation tests and does not explicitly support case/control phenotypes and continuous environments (Lin et al., 2016; Broadaway et al., 2015; Zhao et al., 2015).

4.2 Simulation study

First, we assessed statistical calibration and power of the proposed iSet method using simulated data. We considered the synthetic cohort of 1,000 individuals simulated from genotypes of European populations in the 1000 Genomes Project as described in Section 3.2.1. We considered genetic variants with a minor allele frequency of at least 2%.

4.2.1 Phenotype simulation strategy

Similar to the simulation procedure described in Section 3.2.2, we generated trait measurements in two contexts as the sum of the contribution from a randomly-selected causal region of 30kb ($\boldsymbol{S}$), polygenic background effect ($\boldsymbol{G}$), effects from $K = 10$ unobserved covariates ($\boldsymbol{H}$) and iid noise ($\boldsymbol{\Psi}$):

$$\boldsymbol{Y} = \boldsymbol{S} + \boldsymbol{G} + \boldsymbol{H} + \boldsymbol{\Psi}. \tag{4.25}$$

Genetic effects from the causal region were simulated to generate persistent, rescaling-GxC or general-GxC effects (rescaling-GxC effects + heterogeneity-GxC effects).

- **Persistent and rescaling-GxC effects.** S_r causal variants were randomly selected from the region. Denoting with $\boldsymbol{G} \in \mathbb{R}^{N \times S_r}$ the standardised genotypes of the selected causal variants, the local polygenic effect was simulated as

$$\boldsymbol{S} = \boldsymbol{G}\boldsymbol{b}\,[1, \eta]\,, \quad \text{where } \boldsymbol{b} \overset{\text{iid}}{\sim} \{-1, +1\}\,, \tag{4.26}$$

  where $\boldsymbol{b} \in \mathbb{R}^{S_r}$ is the effect size of the causal variants and η is the proportionality factor of the effect sizes across the two contexts. Note that η can be positive (positive rescaling), negative (negative rescaling) or zero (i.e. the polygenic effect is specific to the first context). Additionally, $\eta = 1$ corresponds to the case of persistent effects.

- **General-GxC.** When simulating general-GxC, we included scenarios with different causal variants between the two contexts. To do so, we independently sampled s_r causal variants in each of the two contexts, resulting in a total of $S_r = 2s_r$ variants. In each context, the causal variants were selected such that all

pairwise squared Pearson correlations were lower than 0.4. This was done by rejecting sampled configurations that did not satisfy this condition. Denoting with $\boldsymbol{G}_1 \in \mathbb{R}^{N \times s_C}$ and $\boldsymbol{G}_2 \in \mathbb{R}^{N \times s_C}$ the standardised genotypes of the casual variants in the two contexts and with $\boldsymbol{b}_1 \in \mathbb{R}^{s_C}$ and $\boldsymbol{b}_2 \in \mathbb{R}^{s_C}$ the respective vectors of their effect sizes, the local polygenetic effect was generated as

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{G}_1 & \boldsymbol{G}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{0}_{s_r \times 1} \\ \boldsymbol{0}_{s_r \times 1} & \boldsymbol{b}_2 \end{bmatrix}, \quad \text{where } \boldsymbol{b}_1, \boldsymbol{b}_2 \overset{\text{iid}}{\sim} \{-1, +1\}. \tag{4.27}$$

The extent of heterogeneity-GxC was controlled by additionally controlling the correlation of the polygenetic effects between the two contexts. In particular, given a certain target correlation range $[\rho_m, \rho_M]$, we only considered realisations for which

$$\rho_m < \text{corr}(\boldsymbol{S}_{:,1}, \boldsymbol{S}_{:,2}) < \rho_M, \tag{4.28}$$

by rejecting realisations that did not satisfy this condition.

The genetic contributions from the regions were scaled so that $v_{\text{r}} = \text{var}\left[\text{vec}(\boldsymbol{R})\right] = 2\%$. When considering the general-GxC case, to limit the extent of rescaling-GxC, we rescaled each column of $\boldsymbol{R}$ to have variance 2%.

Similar to the simulation strategy described in Section 3.2.2, the effects from population structure and unobserved covariates were generated as the sum of a shared and an independent component across contexts.

$$\begin{aligned}
\boldsymbol{G} &= \boldsymbol{G}^{(\text{s})} + \boldsymbol{G}^{(\text{i})} && (4.29)\\
\boldsymbol{H} &= \boldsymbol{H}^{(\text{s})} + \boldsymbol{H}^{(\text{i})} && (4.30)\\
\boldsymbol{G}^{(\text{s})} &\sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{R}, \boldsymbol{a}_{\text{G}}\boldsymbol{a}_{\text{G}}^{\top}\right) && (4.31)\\
\boldsymbol{G}^{(\text{i})} &\sim \text{MVN}\left(\boldsymbol{0}, \boldsymbol{R}, \text{diag}(\boldsymbol{c}_{\text{G}}^2)\right) && (4.32)\\
\boldsymbol{H}^{(\text{s})} &\sim \text{MVN}\left(\boldsymbol{M}\boldsymbol{M}^{\top}, \boldsymbol{a}_{\text{H}}\boldsymbol{a}_{\text{H}}^{\top}\right) && (4.33)\\
\boldsymbol{H}^{(\text{i})} &\sim \text{MVN}\left(\boldsymbol{M}\boldsymbol{M}^{\top}, \text{diag}(\boldsymbol{c}_{\text{H}}^2)\right) && (4.34)
\end{aligned}$$

where $\boldsymbol{M} \in \mathbb{R}^{N \times K}$ is the design matrix of the hidden confounders, $\boldsymbol{R}$ denotes the global realised relatedness matrix and

$$\begin{aligned}
\boldsymbol{a}_{\text{G}} = \sqrt{\boldsymbol{\alpha}_{\text{G}}}, \quad \boldsymbol{c}_{\text{G}} = \sqrt{\boldsymbol{\gamma}_{\text{G}}}, \quad \boldsymbol{a}_{\text{H}} = \sqrt{\boldsymbol{\alpha}_{\text{H}}}, \quad \boldsymbol{c}_{\text{H}} = \sqrt{\boldsymbol{\gamma}_{\text{H}}} && (4.35)\\
\boldsymbol{\alpha}_{\text{G}}, \boldsymbol{\gamma}_{\text{G}}, \boldsymbol{\alpha}_{\text{H}}, \boldsymbol{\gamma}_{\text{H}} \sim \text{Uniform}(0, 1) && (4.36)\\
\boldsymbol{M}_{i,j} \sim \mathcal{N}(0, 1) \quad i = 1, \ldots, N, \quad k = 1, \ldots, K. && (4.37)
\end{aligned}$$

Denoting with α the fraction of shared signal and with β the fraction of residual variance that is non-iid, the different contributions were rescaled such that

$$\begin{aligned}
\text{var}\left[\text{vec}(\boldsymbol{G}^{(\text{s})})\right] &= \alpha v_{\text{bg}} && (4.38)\\
\text{var}\left[\text{vec}(\boldsymbol{G}^{(\text{i})})\right] &= (1-\alpha) v_{\text{bg}} && (4.39)\\
\text{var}\left[\text{vec}(\boldsymbol{H}^{(\text{s})})\right] &= \alpha\beta(1-v_{\text{bg}}-v_{\text{r}}) && (4.40)\\
\text{var}\left[\text{vec}(\boldsymbol{H}^{(\text{s})})\right] &= (1-\alpha)\beta(1-v_{\text{bg}}-v_{\text{r}}) && (4.41)\\
\text{var}\left[\text{vec}(\boldsymbol{\Psi})\right] &= (1-\beta)(1-v_{\text{bg}}-v_{\text{r}}) && (4.42)
\end{aligned}$$

Unless specified otherwise, we considered the default values $v_{\text{r}} = 2\%$, $v_{\text{bg}} = 40\%$, $\alpha = 0.6$ and $\beta = 0.5$. When simulating scenarios with rescaling-GxC we varied the number of causal variants (S_r) and proportionality factor of the effects in the two contexts (η, which quantifies the extent of rescaling). When simulating general-GxC (rescaling-GxC + heterogeneity-GxC) we varied the number of causal variants (S_r) and the range of correlations between the genetic effects in the two contexts ($[\rho_m, \rho_M]$), thereby varying the extent of heterogeneity-GxC. A summary of the parameter values used in simulations is provided in **Table C.2**.

4.2.2 Illustration case

To illustrate the ability of the proposed set tests to detect and characterise associated loci, we considered a sliding-window analysis using simulated data (**Fig. 4.3**). Genetic effects for a quantitative trait in two contexts were generated from three non-overlapping causal regions (each with size 30 kb) selected from a 5 Mb genomic portion on chromosome 13. In particular, we simulated a region with persistent effects, a region with rescaling-GxC effects and a region with heterogeneity-GxC effects (**Fig. 4.3a-b**). The variance explained by of the region was set to 5%. We then applied mtSet, iSet and iSet-het considering a sliding-window analysis in the 5 Mb region (30kb windows, 15kb step). The three tests accurately resolved the genetic architectures at the three causal loci (**Fig. 4.3d-e**). Indeed, the association test (mtSet) detected genetic effects in all three regions, the interaction test (iSet) revealed GxC interactions in the second and the third regions while the test for heterogeneity-GxC (iSet-het) only detected the third region. For comparison, we also considered a conventional multi-trait linear mixed model (Korte et al., 2012) to test for association (mtLMM) and GxC interactions (mtLMM-int), which could not distinguish between the two architectures of GxC interactions (**Fig. 4.3c**).

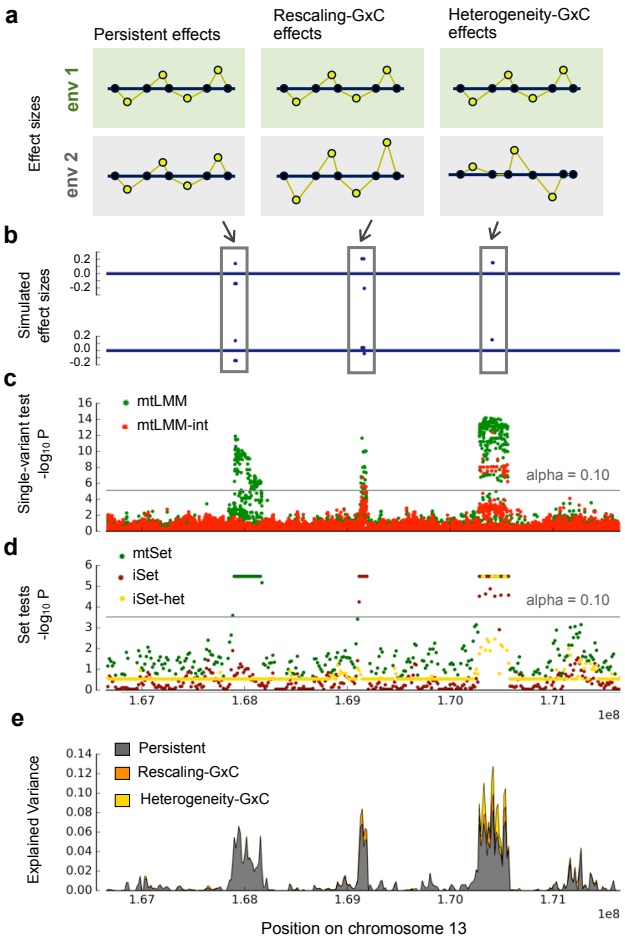


Figure 4.3: **Illustration of iSet using simulated data.** We simulated trait measurements in two contexts with genetic effects from three distinct polygenic loci. (**a**) Schematic representation of the architecture of the three simulated regions. (**b**) Simulated effect sizes on the trait in two contexts. (**c**) Manhattan plots for results from a single-variant LMM (Korte et al., 2012) to test for associations (mtLMM) or interactions (mtLMM-int). Horizontal lines correspond to the $\alpha = 0.10$ significance threshold (Bonferroni adjusted). (**d**) Manhattan plots for results from set test methods, testing for associations (mtSet), interaction effects (iSet) and heterogeneity-GxC effects (iSet-het) in consecutive sets (30 kb regions; step size 15 kb). P values of set tests are bound ($> 10^{-6}$) by the number of null LLR samples used to estimate significance levels. (**e**) The proportion of variance that could be attributed to persistent effects, rescaling-GxC and heterogeneity-GxC, considering the same regions as in (**c-d**).

4.2.3 Calibration of P values

We assessed the calibration of the P values obtained from iSet and iSet-het by simulating only persistent polygenetic effects (no interactions). This analysis was carried out on 100,000 randomly selected 30kb regions. For each region, we generated phenotype values across two contexts, simulating persistent effects from four causal variants and tested for interaction (iSet) and heterogeneity-GxC (iSet-het). P values were estimated using 30 parametric bootstraps for each region/test, resulting in a total of 3,000,000 null LLRs to estimate empirical P values. The QQ plots for iSet and iSet-het of the obtained P values are shown in **Fig. 4.4a**.

Analogously, we assessed the calibration of iSet-het when simulating rescaling-GxC effects. To do so, we pooled results from iSet-het across all the simulations with pure rescaling-GxC that were considered in power simulations (see next section). The resulting QQ plot is shown in figure **Fig. C.1**.

4.2.4 Power comparison

To assess power of iSet for alternative local genetic architectures, we simulated interaction effects in a 30kb region either considering rescaling-GxC effects or general-GxC effects. We compared iSet to the single-variant interaction tests discussed in Section 2.4.3 (mtLMM-SV-int, Korte et al., 2012) using an implementation in LIMIX (see Section 5.2.2, Lippert et al., 2014c). To obtain region-based P values, we used the minimum P value across all the variants in the region, following adjustment for multiple testing. We considered two alternative strategies for this adjustment, i) a conservative Bonferroni approach and ii) eigenMT (Davis et al., 2016), which estimates the number of effect independent tests based on the local linkage disequilibrium (LD, see Section 2.2.2). **Fig. C.2** shows a comparison between the number of variants and the number of effective tests estimated from eigenMT, considering 10,000 30kb-sized regions used in the power simulations (the same comparison is shown also for the other two genotype data considered in this chapter). Note that existing set-based interaction tests cannot be applied to complete designs and hence were not considered (see Section 4.1.4). To obtain P values for iSet and iSet-het we considered 30 parametric bootstraps for each region and computed empirical P values from the 30,000 null LLRs generated in each simulated scenario. For each parameter setting, we considered 1,000 repeat experiments. We used the Benjamini-Hochberg (BH) procedure to adjust for multiple testing across repeat experiments for each method, and assessed different methods in terms of the statistical power at FDR $< 5\%$.

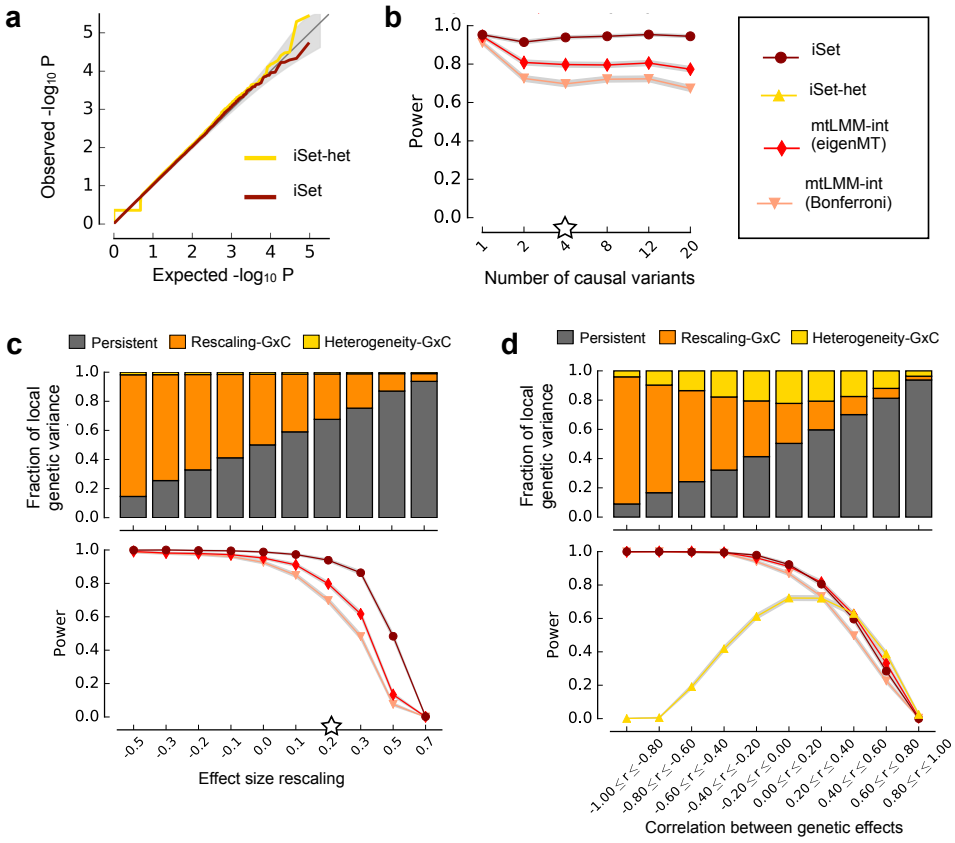


Figure 4.4: **Simulated data to assess power and calibration of iSet.** (**a**) QQ plot for the P values obtained by iSet and iSet-het when only persistent genetic effects were simulated. (**b**) Comparison of the power for detecting genetic interactions on simulated data with rescaling- GxC effects (without heterogeneity-GxC) and for increasing numbers of simulated causal variants. Considered were iSet and a single-variant interaction test (mtLMM-int), using two alternative approaches to adjust for multiple testing of single variants (Bonferroni or eigenMT). (**c**) Lower panel: power comparison as in (**b**), varying the factor of proportionality of the variant effect sizes between contexts; top panel: average fraction of genetic variance explainable by persistent, rescaling-GxC and heterogeneity-GxC effects. (**d**) Analogous comparison as in (**c**) when simulating heterogeneity-GxC effects and varying the correlation of the total genetic effect between contexts. Additionally, iSet-het was used to test for heterogeneity-GxC, which was best powered for large simulated heterogeneity-GxC effects (i.e. low correlation of context-specific genetic signals). White stars indicate default parameter values that were kept constant when varying other parameters for simulated rescaling-GxC (**Table C.2**).

Scenarios with rescaling-GxC. Initially, we considered scenarios with only rescaling-GxC. Varying the number of causal variants in the region (from 1 to 20; **Fig. 4.4b**), we found that iSet was better powered than single-variant models when multiple causal variants were simulated (**Fig. 4.4b**). We then varied the extent of rescaling-GxE effects, which can be controlled using the proportionality factor of the effect sizes between contexts (η, see Section 4.2.1). Negative rescaling ($\eta < 0$, negative x-axis in **Fig. 4.4c**) corresponds to opposite directions of the polygenic effects in the two contexts while $\eta = 0$ corresponds to an effect that is specific to one of the two contexts. The limit $\eta \to 1$ corresponds to the setting with no rescaling, which is equivalent to a scenario of persistent effects across contexts. As expected, all models were best powered for larger rescaling constants and for effects with opposite directions (**Fig. 4.4c**).

Scenarios with heterogeneity-GxC. We then considered simulated scenarios with both rescaling-GxC and heterogeneity-GxC effects and varied i) the extent of simulated heterogeneity-GxC (**Fig. 4.4d**) and ii) the total number of causal variants (**Fig. C.3**). To control the extent of heterogeneity-GxC in **Fig. 4.4d**, we monitored the correlation ρ between the simulated genetic signals from the causal region across the two contexts (see Section 4.2.1). Low to moderate genetic correlations correspond to larger heterogeneity-GxC effects, whereas strongly correlated signals ($\rho \approx \pm 1$) cannot be distinguished from either negative rescaling-GxC effects (for $\rho \approx -1$) or persistent genetic effects (for $\rho \approx 1$). We considered iSet-het to test for heterogeneity-GxC, which was best powered in regimes of low to moderate correlation between genetic effects (power $> 60\%$ for $r^2 < 0.16$, **Fig. 4.4d**). Simulations with larger number of causal variants suggest that iSet robustly detects heterogeneity-GxC **Fig. C.3**.

Comparison with a baseline test for heterogeneity-GxC. To study the accuracy of iSet-het to detect heterogeneity-GxC, we also compared iSet-het with a with a baseline method for detection of heterogeneity-GxC. This method, which we denote as uLMM-het, assigns a heterogeneity-GxC score to genetic regions based on the P values from a univariate LMM and LD information. The scoring strategy is described in the following.

- For regions that fail marginal significance in at least one context, uLMM-het assigns a heterogeneity-GxC score of 0. Marginal significance is assessed using a region-based P value, obtained as the P value of the lead variant from a standard LMM after adjustment for multiple testing using Bonferroni. The intuition behind this scoring rule is that the detection of a significant association in both

contexts is a necessary condition to reveal heterogeneity-GxC effects.

- For regions with significant marginal associations in both contexts, uLMM-het assigns a heterogeneity-GxC score of $1 - R^2$, where R^2 is the squared Pearson correlation between the per-context lead variants.

Intuitively, uLMM-het assigns a high heterogeneity-GxC score only to regions with significant per-context lead variants that are in low LD. For iSet-het, we used the LLR as a score for heterogeneity-GxC.

We used the two methods to score the 20,000 regions considered in the power simulations (**Fig. 4.4c** and **Fig. 4.4d**), half of which harbours heterogeneity-GxC effects. Ranking was assessed using both the receiver-operating characteristic (ROC) and precision-recall (PR) curves (**Fig. C.4**). For uLMM-het, we considered alternative significance thresholds on region-based P values ($P < 0.5, 0.01, 0.01, 0.001$). This experiment confirmed that iSet-het is substantially more accurate for identifying heterogeneity-GxC than this basic univariate approaches.

4.2.5 Variance decomposition

We also investigated the proportion of local genetic variance that can be explained by a model with either persistent, rescaling-GxC or heterogeneity-GxC effects (see Section 4.1.3), when considering different simulated settings (**Fig. 4.4c-d**). A persistent effect model could capture large proportions of the simulated genetic variance, even in the presence of positively correlated GxC, but could not capture variance due to GxC effects with negative rescaling. An LMM that models rescaling-GxC did account for negative and positive rescaling and captured a substantial part of the heterogeneity-GxC effects (**Fig. 4.4d**). Finally, variance contributions that were exclusively captured by a heterogeneity-GxC model were largest for uncorrelated context-specific genetic effects, the same regime where the corresponding test was best powered. We also confirmed that the most general model (i.e., with a full rank trait covariance in the set term) yields unbiased estimates of the total genetic variance in genomic regions, whereas other models yielded biased estimates for some of simulated architectures (**Fig. C.5**).

4.3 Analysis of stimulus-specific eQTLs in monocytes

As a first application to real data, we considered a monocyte stimulus eQTL dataset (Fairfax et al., 2014). In the primary analysis, Fairfax et al. (2014) mapped *cis* and *trans* eQTLs in purified monocytes from 432 healthy Europeans after exposure to interferon-γ

(IFN) for 24 hours and lipopolysaccharide for 2 hours (LPS-2h) or 24 hours (LPS-24h). Mapping genetic effects in stimulated immune cell types is important to characterise the genetic component of common disorders associated with a malfunctioning of immune activity and inflammation, e.g. atherosclerosis, inflammatory bowel disease and cancer (Fairfax et al., 2014). The study revealed master *trans*-eQTL regulators and characterised the molecular mechanism underlying GWAS loci associated with inflammatory diseases. Here, we focus on the local (*cis*) genetic architecture of gene expression and how genetic effects change across different cellular contexts. To do so, we performed a naive/stimulus pairwise analysis for each of the three stimulated states (i.e. naive/IFN, naive/LPS-2h and naive/LPS-24h; see also **Fig. 4.5a**).

4.3.1 Data preprocessing

Gene expression levels in the naive state, and after exposure to IFN, LPS-24h and LPS-2h were available for 414, 367, 322 and 261 individuals respectively. Normalisation, correction for batch and probe filtering were performed by Fairfax et al as described in the primary study. We discarded probes that did not have an associated ENSEMBL ID, resulting in 12,661 probes for analysis (out of 15,421). We further limited analysis to the 288 individuals for which gene expression levels were available in all the four cellular contexts. To account for hidden covariates and confounding factors, we applied PEER (Stegle et al., 2012) with default parameter values, fitting 30 hidden factors across all samples (individuals and stimulus states). PEER residuals for each gene and context were quantile-normalised to a unit variance normal distribution and used for all subsequent analysis.

4.3.2 Mapping of associations and interactions

For each of the 12,661 probes, we used iSet to test for pairwise interaction effects, considering the naive monocyte state and each stimulus condition in turn, performing a single test using proximal (putatively *cis* acting) variants in a 100 kb region centred on the transcription start site. We tested for *cis* associations (mtSet), GxC interactions (iSet) and heterogeneity-GxC (iSet-het). For comparison, we applied single-variant tests in the same regions, testing for association (mtLMM) and GxC interaction (mtLMM-int). For single-variant tests, we estimated gene-level significance using the P value of the lead *cis* variant adjusted within *cis* regions using eigenMT, Brown et al., 2016. Empirical P values for mtSet, iSet and iSet-het were estimated from 30 permutations/parametric bootstraps per-gene and stimulus, combining all null LLRs across genes

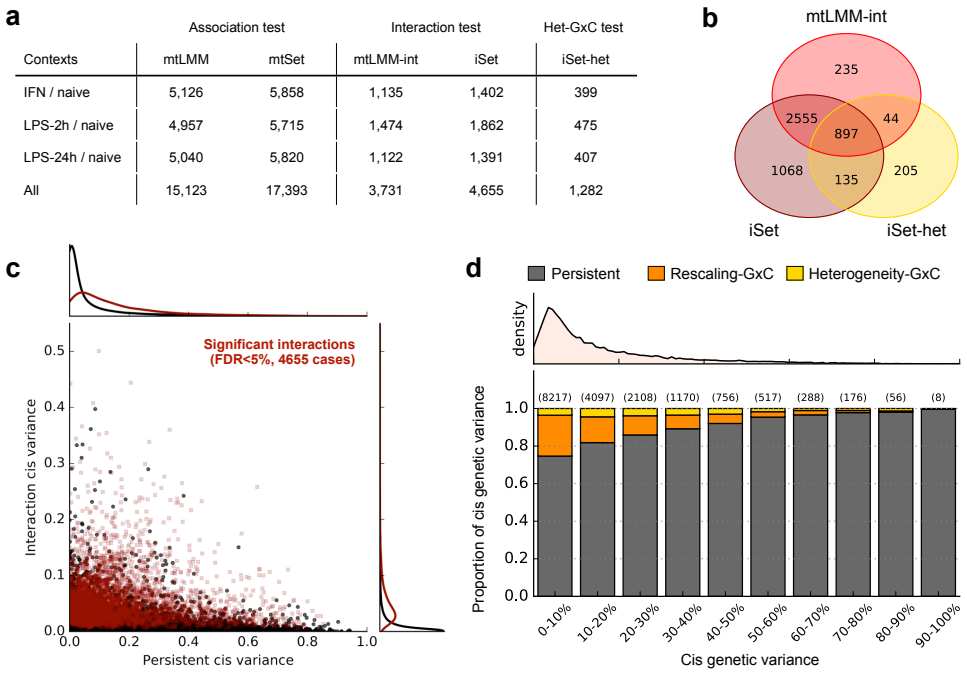


Figure 4.5: **Analysis of stimulus-specific eQTLs in monocytes.** (**a**) Number of probes with at least one significant *cis* association (Association test) or genotype-stimulus interaction (Interaction test) for alternative methods and stimulus contexts. Considered were the proposed set tests (mtSet, iSet, iSet-het) as well as single-variant multi-trait LMMs (mtLMM, mtLMM-int), testing for genetic effects in *cis* (100kb region centred on the transcription start site; FDR $< 5\%$). Additionally, iSet-het was used to test for heterogeneity-GxC effects. Individual rows correspond to different stimulus contexts with "All" denoting the total number of significant effects across all pairwise analyses. (**b**) Venn diagram of significant effects detected by alternative methods and tests (across all pairwise analyses). (**c**) Bivariate plot of the variance attributed to persistent genetic effects versus genotype-stimulus interactions for all probes and stimuli. Significant interactions are shown in red. Density plots along the axes show the marginal distributions of persistent genetic variance (top) and variance due to interaction effects (right), either considering all (black) or the subset of probe/stimulus pairs with significant interactions (red). (**d**) Average proportions of *cis* genetic variance attributed to persistent effects, rescaling effects and heterogeneity-GxC, considering probes/stimuli with significant *cis* effects (5% FDR, mtSet) and stratified by increasing fractions of the total *cis* genetic variance. Shown on top of each bar is the number of instances in each variance bin. The top panel shows density of genes as a function of the total *cis* genetic variance.

(resulting in 379,830 null LLRs per stimulus and statistical test, see Section 4.1.2). Results from all methods were adjusted for multiple testing across genes using the BH procedure applied to each naive/stimulus analysis separately. We defined significant associations, interactions and heterogeneity-GxC effects at genome-wide FDR $< 5\%$ (**Fig. 4.5a**).

Although many of the genes with significant associations from single-variant and set-based tests were consistent (**Fig. 4.5b**), set tests offered increased power for detecting both associations and interactions (24.8% power increase for interactions; 4,655 versus 3,731 probes and stimuli with an interaction; FDR $< 5\%$, **Fig. 4.5a**, **Fig. C.6**). Additionally, iSet-het yielded 1,282 genes with heterogeneity-GxC effects (**Fig. 4.5a-b**). This suggests that GxC effects are frequently associated with different regulatory architectures between contexts, which is consistent with previous reports (Fairfax et al., 2014).

Although on average the proportion of variance explained by GxC tended to be smaller than for persistent effects (median 4.0% for GxC versus median 10.4% for persistent effects, for genes with significant GxC, **Fig. 4.5c**), GxC was the dominant genetic source of variation for 11.7% of the significant *cis* eQTLs (**Fig. 4.5d**; defined as explaining 50% or more of the *cis* genetic variance). Consistent with previous reports (Gagneur et al., 2013; GTEx Consortium, 2015), we observed that genes with large relative GxC effects were associated with weak overall *cis* effects, whereas large-effect eQTLs tended to be persistent across the jointly analysed contexts (**Fig. 4.5d**).

4.3.3 Mechanistic underpinning of heterogeneity eQTLs

Heterogeneity eQTLs correspond to weakly correlated polygenic signal across contexts. We expect that heterogeneity eQTLs are associated to i) marginally significant genetic effects in both the analysed contexts and ii) weakly correlated polygenic signals. To test this hypothesis we applied a univariate set test (stSet, see Section 3.1.1) to each cellular context independently, considering the same *cis* genetic regions. For each gene and naive/stimulus pair,

1. the strength of the marginal associations was quantified as the -$\log_{10}$ of the maximum stSet P value across the jointly analysed contexts (after adjustment for multiple testing across genes using BH);

2. the correlation between polygenic signals was estimated as the correlation between the *cis* Best Linear Unbiased Predictors (BLUP) from stSet[3] across the jointly analysed contexts.

Fig. 4.6 shows a scatter plot across all genes and probe/stimulus pairs of these two quantities. Heterogeneity eQTLs are coloured in red. The results show very clearly that heterogeneity eQLTs correspond to instances where marginal associations are strong and polygenic signals are weakly correlated across contexts, which confirms our hypothesis.

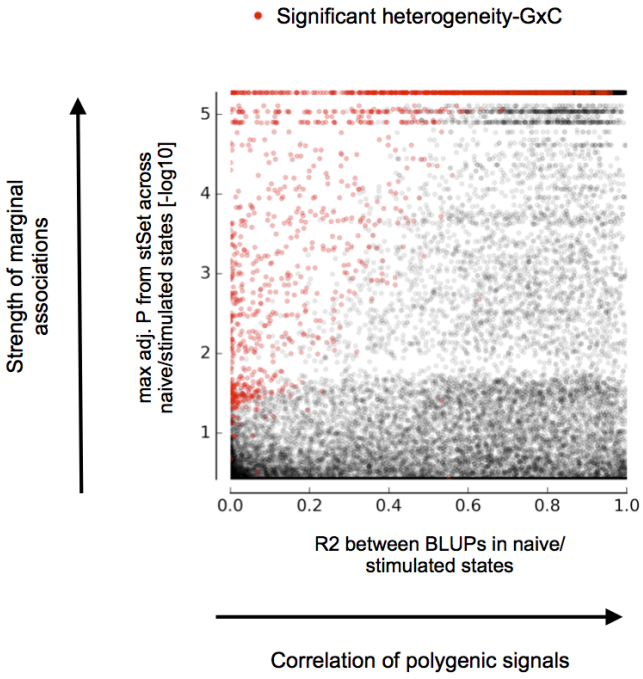


Figure 4.6: **Heterogeneity eQTLs correspond to significant marginal associations and weakly correlated polygenic signals.** Scatter plot across all probe/stimulus pairs of the maximum adjusted P value between the naive state the corresponding stimulated state versus the correlation of the BLUP across the two contexts. Heterogeneity eQTLs are highlighted in red. The figure clearly shows that heterogeneity eQTLs are association to strong marginal associations with weakly correlated polygenic signals between the two contexts.

[3]Considering the stSet model in Eq (3.4), the BLUP for the set component is

$$\boldsymbol{y}_{\star}^{(\text{set})} = \hat{\sigma}_r^2 \boldsymbol{R}_r \left(\hat{\sigma}_r^2 \boldsymbol{R}_r + \hat{\sigma}_g^2 \boldsymbol{R}_g + \hat{\sigma}_n^2 \boldsymbol{I}_N\right)^{-1} (\boldsymbol{y} - \boldsymbol{F}\hat{\boldsymbol{b}}),$$

where $\hat{\cdot}$ denotes the restricted MLE of the corresponding parameter. See also description in Section 2.3.6.

Heterogeneity eQTLs are associated with complex genetic architectures. Next, we assessed whether classes of genes with significant associations, interactions and heterogeneity-GxC effects were associated with different complexities of the genetic architecture, i.e. different numbers of independent associations. To estimate the number of independent genetic effects (up to three), we applied a single-variant step-wise selection LMM (Segura et al., 2012) to each gene and context independently. Briefly, we considered iterative analyses in the region of interest, where in each step the lead variant from the previous step is added in the model as an additional fixed effect covariate. For each context, region-based P values were adjusted for multiple testing across genes using the BH procedure for each of the three steps. We assessed marginal significance and counted the number of signals at FDR$<$0.05. This analysis yielded 12,371, 1,646 and 123 instances (across all genes and contexts) with one, two or three associations respectively.

Fig. 4.7 shows the cumulative fraction of different classes of probe/stimulus pairs (all genes and those with association, interaction and heterogeneity-GxC) with the average numbers of independent associations across the jointly analysed contexts. As expected, genes with significant heterogeneity-GxC were more likely to harbour multiple independent associations, confirming that heterogeneity-GxC eQTLs have complex genetic architectures.

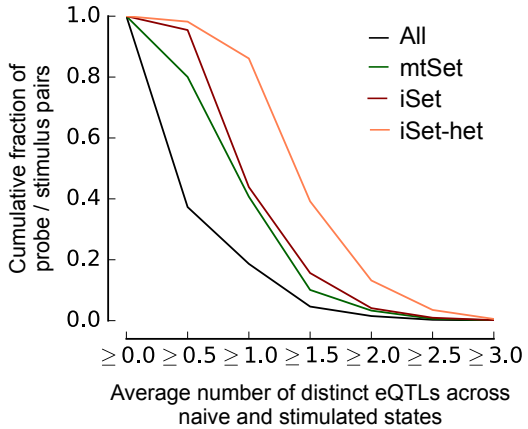


Figure 4.7: **Cumulative fraction of gene/stimulus pairs with increasing numbers of distinct univariate eQTLs for different gene sets.** Shown are the fractions for all probe/stimulus pairs (All), probe/stimulus pairs with significant *cis* associations (mtSet), probe/stimulus pairs with significant GxC (iSet) and instances with significant heterogeneity GxC (iSet-het) with the average numbers of distinct univariate eQTLs from a step-wise selection across the corresponding naive/stimulated state.

Breakdown of heterogeneity eQTLs into different classes. We then used a simple method based on the summary statistics from the step-wise analysis and LD information to define different classes of context-specificity and annotate heterogeneity eQTLs.

For each gene/stimulus pair, cases with significant marginal associations in both contexts and context-specific lead variants in low LD ($R^2 < 0.20$) were associated to a change in the lead variant (**Fig. 4.8a**). For cases with significant marginal associations in both contexts and a shared lead eQTL (lead variants $R^2 > 0.80$), we annotated context-specific secondary effects when either i) the secondary effect was significant in only one of the two contexts or ii) the context-specific lead variants of the secondary signal were in low LD ($R^2 < 0.20$). We find 1,218 gene/stimulus pairs associated with a shift in the lead variant and 3,874 gene/stimulus pairs associated with shared effects, with 700 cases having context-specific secondary eQTLs.

Overlaying heterogeneity eQTLs with these different classes, we attributed 42% of the heterogeneity-GxC effects (542 out of 1,282) to context-specific lead variants (**Fig. 4.8a-b**). For an additional 13% of the heterogeneity eQTLs (165 out of 1,282) we found context-specific secondary effects (**Fig. 4.8a,c**). Interestingly, the remaining 574 heterogeneity eQTLs (45%) could not be annotated using the considered single-variant approach. For 46% of the heterogeneity-GxC cases without clear single-variant interpretation (266 out of 574), the lead variants from the single-variant LMM were marginally significant but in weak linkage ($0.2 < R^2 < 0.8$, see example in **Fig. 4.8d**). These were also cases where it is hard to assess the presence of independent genetic signals using exclusively single-variant methods. The annotation of heterogeneity-GxC eQTLs may also be hampered by the lower power of the univariate approach in comparison with iSet, which integrates across multiple variants and contexts (Flutre et al., 2013). Indeed, we find that for another 30% of the instances without annotation (176 out of 574) the single-variant method did not detect a significant association in at least one of two jointly analysed contexts (**Fig. C.7**).

4.3.4 Note on opposite effects

Finally, we tested whether polygenic signals with opposite effects across contexts are associated with heterogeneity-GxC. We classified the 3,874 instances with significant marginal associations in both contexts and a shared lead eQTL (lead variants $R^2 > 0.80$) into same-direction and opposite-direction eQTLs. The classification was based on the sign of the correlation of the genetic effects across contexts, defined as sign $(\beta_1 \beta_2 \text{corr}\,(\boldsymbol{g}_1, \boldsymbol{g}_2))$ where $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are the genotypes of the lead-variants and β_1 and β_2 their effect sizes

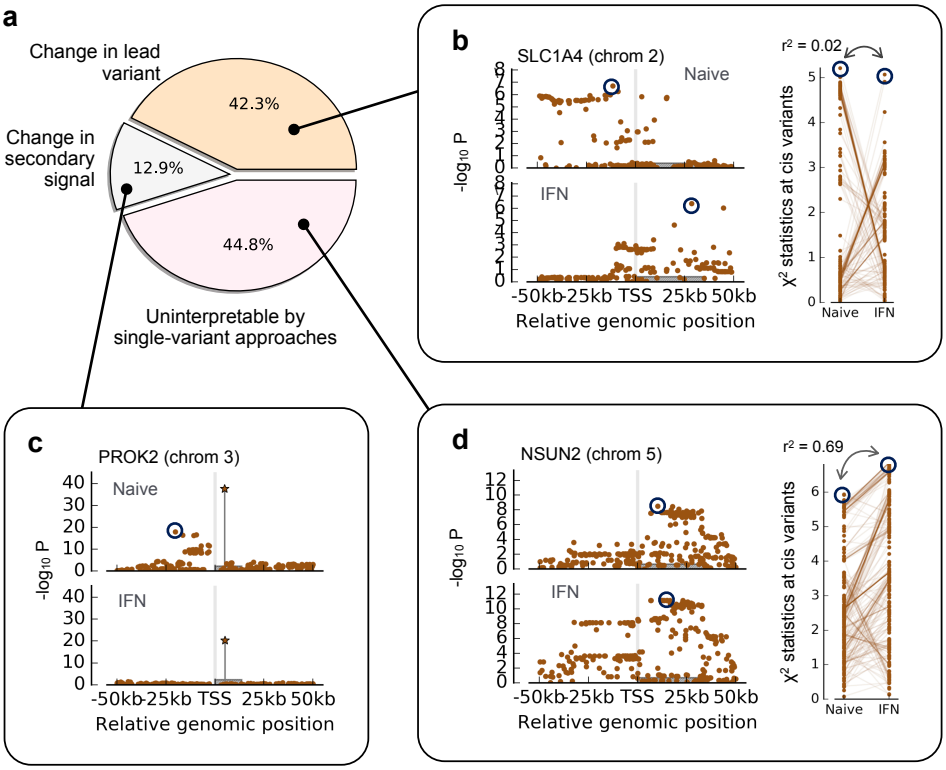


Figure 4.8: **Characterisation of genes with significant heterogeneity GxC for stimulus eQTLs in monocytes.** (**a**) Breakdown of 1,282 probe/stimulus pairs with significant heterogeneity GxC into different classes using a single-variant step-wise LMM. (**b-d**) Manhattan plots for representative genes with significant heterogeneity GxC. The gene body is shown as grey box. (**b**) Manhattan plot (left) and r^2 statistics for variants in both contexts (right) for the gene *SLC1A4*. Dark circles indicate distinct lead variants in both contexts (R^2<0.2). (**c**) Manhattan plot from the second step of the step-wise selection analysis for the gene *PROK2*. The start symbol indicates the shared lead variant in both contexts. The single- variant analysis reveals a secondary association signal that is specific to the naive state. (**d**) Analogous plot as in c for the gene *NSUN2*, which could not be interpreted using single- variant analyses.

respectively. These criteria led to the identification of 53 opposite-direction eQTLs and 3,821 same-direction eQTLs across all naive/stimulus analysis. Notably, iSet-het detected significant heterogeneity-GxC for 11 of the opposite-direction QTLs, a 3.6 fold enrichment ($P < 10^{-3}$, one-sided Fisher's exact test) compared to eQTLs with consistent effect directions between contexts (261 gene/stimulus pairs with significant heterogeneity-GxC out of 3,821 eQTLs with consistent direction). There is a concern that the heterogeneity-GxC tests are not independent as a consequence of the fact that

the naive state is considered in all pairwise analyses. To address this we repeated the enrichment analyses for each naive/stimulus pair independently. Results are summarised in **Table C.3**. For naive/IFN, naive/LPS2 and naive/LPS24 we find a fold enrichment of 5.2 ($P < 0.05$), 1.11 ($P = 0.56$) and 7.2 ($P < 10^{-3}$). Taken together, these results suggest that changes in the configuration of causal variants may underly a substantial fraction of seemingly opposite-direction *cis* eQTLs, a result that is consistent with recent findings (GTEx Consortium, 2015). Interestingly, among the genes that were identified as opposite-effect eQTLs in the primary analysis of the same data (Fairfax et al., 2014) we found three cases with strong heterogeneity-GxC signal (*OAS1*, *LMNA* and *PTK2B*; see **Fig. C.8**).

4.4 Extension to analysis of stratified designs

Thus far, we have focused on settings with repeat measurements, where the same trait is observed in all individuals and contexts. I here discuss applications of iSet to studies where each individual is phenotyped in only one context. This scenario occurs when stratifying a population using an external context variable (see **Fig. 4.9**).

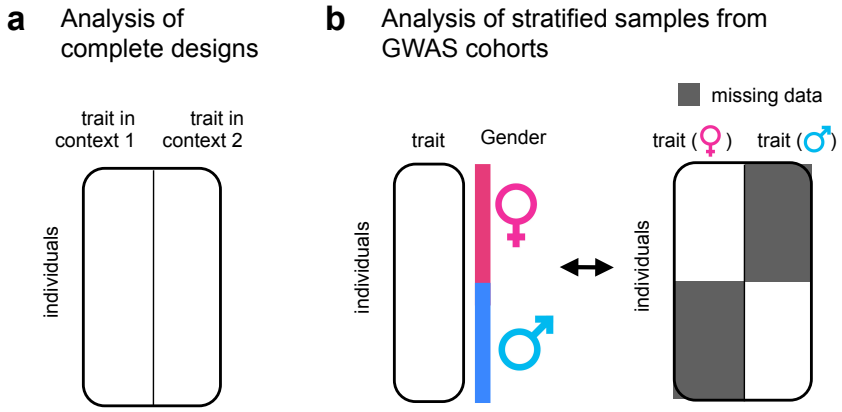


Figure 4.9: **Alternative study designs supported by iSet.** iSet supports efficient interaction set tests both for datasets with complete designs (where each individual is phenotyped in all analysed contexts, (**a**), and stratification analyses (where each individual is phenotyped in only one of the analysed contexts, (**b**).

After deriving the model in Section 4.4.1, I discuss validation in simulations in Section 4.4.2. Finally, in Section 4.4.3 I discuss application to a gene-by-sex interaction analysis of lipid traits in the Northen Finland Birth Cohort (NFBC) (Sabatti et al., 2009).

4.4.1 Model derivation

Let us consider a dataset with N unrelated individuals in $C = 2$ contexts. If trait observations are available in all individuals and contexts we can consider the model

$$\operatorname{vec}(\boldsymbol{Y}) \sim \mathcal{N}\left(\operatorname{vec}(\boldsymbol{FB}), \boldsymbol{C}_r \otimes \boldsymbol{GG}^\top + \boldsymbol{C}_n \otimes \boldsymbol{I}\right) \tag{4.43}$$

where $\boldsymbol{Y}$ is the $N \times C$ phenotype matrix, $\boldsymbol{F} \in \mathbb{R}^{N \times K}$ the design matrix of K fixed effect covariates, $\boldsymbol{B} \in \mathbb{R}^{K \times C}$ their effect sizes and $\boldsymbol{G}$ the $N \times R$ standardised genotype matrix for R variants and $\boldsymbol{C}_r$ and $\boldsymbol{C}_n$ are $C \times C$ covariance matrices. Note that in Eq (4.43) we did not consider a relatedness component. Similar to the mtSet-PC model (see Section 3.1.4), population structure can be accounted for by introducing the first principal components of the RRM as fixed effect covariates. This assumption is necessary to derive an efficient inference scheme for stratification analyses.

Let us consider the case where the first $N_1 < N$ individuals have been phenotyped in one of the contexts while the remaining $N_2 = N - N_1$ have been phenotyped in the other context. It is convenient to introduce the extended $N \times N$ context covariance matrices $\tilde{\boldsymbol{C}}_r$ and $\tilde{\boldsymbol{C}}_n$ for the set and noise component respectively as

$$\tilde{\boldsymbol{C}}_{r\, n_1,n_2} = \boldsymbol{C}_{r\, e(n_1),e(n_2)} \tag{4.44}$$

$$\tilde{\boldsymbol{C}}_{n\, n_1,n_2} = \boldsymbol{C}_{n\, e(n_1),e(n_2)}\,, \tag{4.45}$$

where $e:\ n \in [1,N] \longrightarrow e(n) \in \{1,2\}$ is an indicator function that returns the context $e(n)$ in which sample n has been observed. For convenience, we also introduce

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{F}_{1:N_1,:} & \boldsymbol{0}_{N_1 \times K} \\ \boldsymbol{0}_{N_2 \times K} & \boldsymbol{F}_{N_1+1:N,:} \end{bmatrix} \in \mathbb{R}^{N \times 2K}. \tag{4.46}$$

Denoting with $\boldsymbol{y} \in \mathbb{R}^N$ the vector of the observed trait measurements, marginalising out the unobserved trait measurements in the model in Eq (4.43) results in

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{Xb}, \tilde{\boldsymbol{C}}_r \odot \boldsymbol{GG}^\top + \tilde{\boldsymbol{C}}_n \odot \boldsymbol{I}\right), \tag{4.47}$$

where we set $\boldsymbol{b} = \operatorname{vec}(\boldsymbol{B})$. This model has two important properties:

1. The first covariance term can be written as a low-rank covariance

$$\tilde{\boldsymbol{C}}_r \odot \boldsymbol{GG}^\top = \boldsymbol{WW}^\top \quad \text{where} \quad \boldsymbol{W} = \left(\left[\boldsymbol{C}_r^{1/2}\right]_{e,:} \otimes \boldsymbol{1}_{1\times R}\right) \odot (\boldsymbol{1}_{1\times 2} \otimes \boldsymbol{G}) \tag{4.48}$$

where $\boldsymbol{e}_i = e(i)$ $i = 1, \dots N$, and

$$\text{rank}(\tilde{\boldsymbol{C}}_r \odot \boldsymbol{G}\boldsymbol{G}^\top) = \text{rank}(\boldsymbol{C}_r) \cdot R, \tag{4.49}$$

assuming $\text{rank}(\boldsymbol{C}_r)R < N$.

2. The likelihood does not depend on the off-diagonal elements of $\tilde{\boldsymbol{C}}_n$[4] and hence without loss of generality, we can consider a diagonal matrix for the noise covariance.

Thus, the model in Eq (4.47) can be expressed as

$$\boldsymbol{y} \sim \mathcal{N}\left(\underbrace{\boldsymbol{X}\boldsymbol{b}}_{\text{fixed effect covariates}}, \underbrace{\boldsymbol{W}\boldsymbol{W}^\top}_{\text{low-rank set component}} + \underbrace{\boldsymbol{D}}_{\text{diagonal noise component}}\right), \tag{4.50}$$

where

$$\boldsymbol{D} = \text{diag}(\underbrace{\sigma_1^2, \dots, \sigma_1^2}_{N_1}, \underbrace{\sigma_2^2, \dots, \sigma_2^2}_{N-N_1}) \tag{4.51}$$

with σ_1^2 and σ_2^2 denoting the context-specific noise variances. Note that in the model in Eq (4.50) the total covariance is the sum of a low-rank and a diagonal matrix. This allows us us to reduce the naive $O(N^3)$ complexity to $O(NR^2 + NK^2 + R^3 + K^3)$ as shown in Section A.6.

Single-variant models for stratification analysis. For comparison, we considered a single-variant interaction test with context-specific noise levels

$$\tilde{\boldsymbol{y}} \sim \mathcal{N}\left(\underbrace{\boldsymbol{X}\boldsymbol{b}}_{\text{fixed effect covariates}} + \underbrace{\boldsymbol{e}\alpha}_{\text{context}} + \underbrace{\boldsymbol{g}\beta}_{\text{genetic variant}} + \underbrace{(\boldsymbol{g} \odot \boldsymbol{e})\,\gamma}_{\text{GxC}}, \underbrace{\boldsymbol{D}}_{\text{heter. noise}}\right) \tag{4.52}$$

where $\boldsymbol{g} \in \mathbb{R}^N$ denotes the genotype vector of the variant to be tested, $\boldsymbol{e}$ is a binary indicator specifying in which context the phenotype has been observed, α is the fixed effect of the context variable, β the fixed effect of the genetic variant and γ the fixed effects of their interaction. The test for interaction corresponds to testing for $\gamma \neq 0$. Following Korte et al. (2012) and Furlotte and Eskin (2015), the total covariance matrix

[4]This can be easily understood from the fact that in a stratified design there are no repeat trait measurements of the same individuals across contexts so that the residuals correlation across contexts cannot be estimated.

is estimated under the no-association model while for single-variant testing only the total variance is updated (see also Section 2.4.4).

4.4.2 Simulations

To study the performance of iSet when considering interaction analyses of stratified populations, we carried out simulation experiments analogous to those considered for fully observed designs (Section 4.4). We generated a synthetic cohort of 2,000 Europeans using the strategy described in Section 3.2.1 (dataset *simPopStruct*). The context in which each phenotype was observed was independently sampled from a Bernoulli distribution with rate 50%. Population structure was accounted for including the first ten principal components of the RRM as fixed effect covariates. We did not consider tests for heterogeneity-GxC, as differential tagging of causal variants could potentially result in spurious heterogeneity-GxC signals, and hence additional controls would be required.

We compared iSet to the single-variant interaction test for incomplete designs introduced in the previous section and the gene-environment set association test described in Section 4.1.4 (GESAT, Lin et al., 2013). The latter approach is representative for a family of very similar set tests that can only be applied to test for interaction effects in stratified populations. GESAT was run using the function GESAT of the package iSKAT version 1.2. Both iSet and GESAT were applied on identically processed standardised variants. Results are shown in **Fig. 4.10**.

We again confirmed statistical calibration of iSet and found similar power benefits as for complete designs. Notably, iSet was consistently better powered than GESAT, most likely because GESAT does not explicitly model correlations of the local genetic effect between contexts (see Section 4.1.4). Surprisingly, the single-variant interaction test performed slightly better than iSet and GESAT in the range of positive correlations when general-GxC was simulated (**Fig. 4.10d**).

4.4.3 Application to gene-by-sex interaction analysis in lipid traits

Next, we applied iSet to a genome-wide analysis of genotype-sex interactions in the same four lipid-related traits we considered in Section 3.3.1 (fasting HDL and LDL cholesterol levels, triglycerides and C-reactive protein). Note that while in Section 3.3.1, we considered a multi-trait analysis across the four traits, here we perform a gene-sex interaction analysis independently for each of these trait. The data were preprocessed as described in Section 3.3.1. Similar to the multi-trait analysis, we tested consecutive

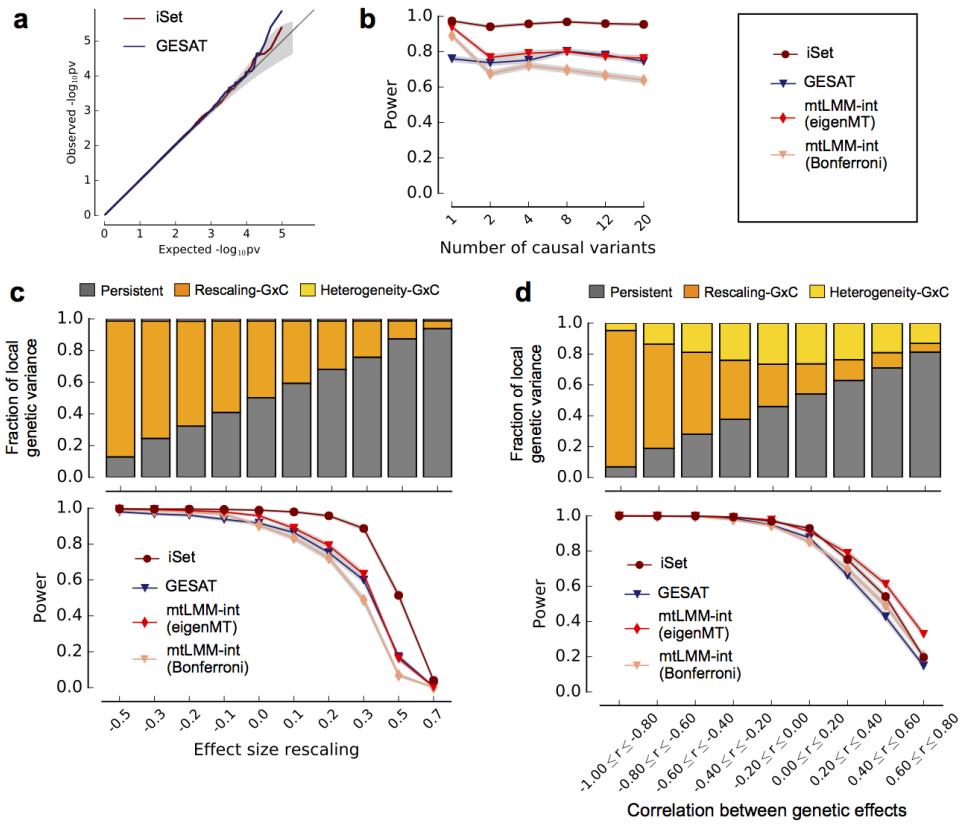


Figure 4.10: **Simulated data to assess power and calibration of iSet for analyses of stratified designs.** (**a**) QQ plot for the P values obtained when applying iSet and GESAT to synthetic datasets where only persistent genetic effects were simulated. (**b-d**) Power comparison of iSet and alternative methods using simulated data where each individual is phenotyped in one of two contexts. Shown are power comparisons for alternative methods, analogous to **Fig. 4.4**. (**b**) Power to detect genetic interactions when simulating rescaling-GxC for increasing numbers of causal variants. (**c**) Power comparison when varying the factor of proportionality of the variant effect sizes between contexts. (**d**) Power comparison when varying the correlation of the simulated genetic effects. We considered iSet, a single-variant interaction test (mtLMM-int, [10]) as well as the interaction sequence kernel association test (GESAT, Lin et al., 2013), a set test designed for stratified populations. For single-variant models, two alternative adjustments for multiple testing were considered (Bonferroni, eigenMT).

100kb regions for genotype-sex interactions with a step size of 50kb. We considered both the association set test (mtSet) and the interaction set test (iSet), which we compared to a single-variant interaction test (mtLMM-int), GESAT and a univariate set test without stratification by sex (stSet). In order to correct for population structure we considered the first ten principal components of the RRM as fixed effect covariates. For each window we considered 100 permutations for mtSet and stSet and 100 parametric bootstraps for iSet. The obtained null LLRs were combined across windows and traits for each method to obtain empirical P values. Significance of the considered statistical tests was assessed at FWER=10%. Manhattan plots and QQ plots for all methods are shown in **Fig. C.9** and **Fig. C.10** respectively.

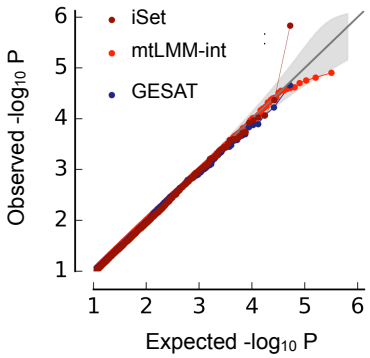


Figure 4.11: **QQ-plot of P values from genotype-sex interaction tests for C-reactive protein levels considering different methods.** QQ-plot of P values from genotype-sex interaction tests for C-reactive protein levels in individuals from the Northern Finland Birth Cohort Study considering iSet, a single-variant interaction test (mtLMM-int) and the interaction sequence kernel association test (GESAT).

iSet retrieved one genome-wide significant interaction effect (C-reactive protein, chr1:40,450,000; $P = 1.47 \cdot 10^{-6}$; FWER $< 10\%$), whereas alternative set tests and single-variant models did not yield any significant effect (**Fig. 4.11**, **Fig. C.9** and **Fig. C.10**). There may be a concern that the reduced power of a single-variant interaction test is due to poor tagging of genetic signals from the sparse genotype in the NFBC1966 cohort. To explore this, we imputed genotype data in the region using the 1000 Genomes Project phase 3 reference panel in a 5Mb region centred in the association found by iSet. After aligning the dataset to the reference panel, we ran shapeit v2.r727 (Delaneau et al., 2014) with the recommended parameters to estimate haplotypes. We then used impute2 v2.3.2 (Howie et al., 2012) with the recommended parameters to impute untyped genotypes. Notably, the single-variant approach mtLMM-int

did not identify significant associations on imputed data (**Fig. C.11**).

The interaction found by iSet on chromosome 1 is within 400 kb from a replicated locus identified in a large meta-analysis (66,185 individuals, $P < 6{\cdot}10^{-11}$ Dehghan et al., 2011). This study also reported a marginally significant interaction effect with sex at the same locus ($P < 5 \cdot 10^{-3}$). This replication is encouraging, however further work is required to assess whether the polygenic locus identified by iSet and the GWAS variant underlie the same causal variant. Finally, a local single-variant analysis, separately for female and male individuals, provided evidence that this interaction most likely reflects a male-specific genetic effect (**Fig. C.12**).

iSet revealed a second suggestive interaction with sex for LDL cholesterol levels (chr3:121,850,000, **Fig. C.9**). Although this effect failed genome-wide significance (FWER<20%, iSet), iSet again provided stronger statistical evidence than other approaches ($P_{\text{iSet}} = 3.7 \cdot 10^{-6}$, $P_{\text{GESAT}} = 4.8 \cdot 10^{-6}$, $P_{\text{mtLMM-int}} = 3.2 \cdot 10^{-5}$). Among the genes at this locus is *ADCY5*, which has been linked to blood glucose levels in large meta-analysis (Dupuis et al., 2010; Saxena et al., 2010) and hence is a plausible candidate to affect LDL via glucose regulation (Otero et al., 2002).

Finally, we note that stratification of quantitative traits by context can also be used to increase power for detecting associations rather than interactions, which is similar to previous strategies applied in the context of single-variant analyses of quantitative (Kim et al., 2014) and categorical traits (Gauderman et al., 2013; Morris et al., 2010). Using this approach, we identified three additional associations when using mtSet on sex-stratified individuals, which were missed by standard set tests and other methods (**Fig. C.9**). These include the same locus with a sex-specific effect on C-reactive protein ($P = 1.42 \cdot 10^{-7}$ using mtSet, $P = 1.89 \cdot 10^{-3}$ using a standard set test) and an association for HDL cholesterol levels and triglycerides, loci that harbour two replicated associations (Sabatti et al., 2009; Kraja et al., 2011).

The interaction and association loci discussed here are highlighted in the Manhattan plots in **Fig. C.9**. A tabular summary is provided in **Table C.4**.

4.5 Summary and discussion

In this chapter, I have presented iSet, a method based on linear mixed-models to test for gene- context interactions using variant sets. On simulated data we have shown that iSet offers power advantages compared to previous methods. Additionally, we have shown that set-based models are effective for disentangling the genetic architecture of interaction effects, differentiating between consistent changes in genetic effects between

contexts and changes in the configuration of causal variants. The underlying test for heterogeneity-GxC we propose is related to co-localization tests (Wallace, 2013; Wallace et al., 2012; Fortune et al., 2015), however with an opposite objective.

In an application to a stimulus eQTL study, we have shown that approximately 25% of the gene-stimulus interactions are associated with significant heterogeneity-GxC. This suggests that complex changes in the genetic architecture between cellular contexts are relatively frequent. Additionally, we have observed that genes with opposite effects are enriched for heterogeneity-GxC. This finding points to a possible bias, whereby opposite effects identified using single-variant models may in part be due to context-specific causal variants that are tagged by a common lead variant.

The proposed iSet model is not free of limitations. First, scalable inference in our model requires low-rank assumptions, meaning that the number of variants in the analysed region is small compared to the number of individuals. Similar to mtSet, there may be trade-offs between the size of variant sets and the number of samples, in particular for densely imputed or sequenced cohorts. The inference scheme we have derived is efficient if phenotypes are either observed in all contexts and individuals or, alternatively, if a cohort has been stratified using a context variable. Intermediate designs can also be handled but currently require the use of separate imputation schemes (Dahl et al., 2016). It is also worth noting that the test for heterogeneity-GxC (iSet-het) will be most accurate if all individuals are phenotyped in each context. Although in principle the model can also be used in stratified designs, there may be concerns that heterogeneity results from technical factors, for example due to differences in genotyping accuracy or variant allele frequencies in the corresponding sub populations.

Finally, although iSet is general, we here have focused on the pairwise analysis of alternative contexts. By jointly analysing multiple, related context and traits, it would be possible to obtain a more comprehensive picture by discerning the regulatory architectures within regions. Extensions of iSet in this direction are an area of future work.

5 | Flexible LInear MIXed models

Although a substantial number of different genetic analyses can be cast within the linear mixed model (LMM) framework, most efficient software implementations are designed for solving a narrow set of tasks. One reason for this is a trade-off between flexibility and computational efficiency, whereby efficient LMM implementations are specialised to exploit algebraic properties that are specific to the considered model. The design of software that enables flexible modelling while retaining computational efficiency whenever possible, is challenging. In this chapter, I discuss LIMIX, a framework for LInear MIXed models that aims at overcoming these challenges. LIMIX builds on a flexible inference framework that allows designing Gaussian models with composite covariance matrices, encoded as functions of model parameters. Building on this core structure, higher-level modules facilitate different single-trait and multi-trait genetic analyses, including variance decomposition, single-variant tests, set-based tests and genomic prediction models. All the models in this thesis, including those that we considered for comparison, are implemented and available in LIMIX (with the exception of the iSKAT model). The development of LIMIX was motivated by several applied projects I have contributed to during my PhD. In these projects, I have designed, implemented and validated customised genetic models to address specific analysis needs. LIMIX has been used to perform different genetic analyses (Dubin et al., 2015a; Sasaki et al., 2015; Kawakatsu et al., 2016; Horton et al., 2016; Sudmant et al., 2015; Märtens et al., 2016; Baud et al., 2017; Schor et al., 2017; Cannavò et al., 2016; Chen et al., 2016).

In Section 5.1, I discuss the core flexible framework for parameter inference. In Section 5.2, I describe two high-level modules for specific genetic analyses: a module for variance component analysis and a module for single-variant association testing for the analysis of multiple traits. To demonstrate the utility of this software suite, in Section 5.3, I describe two analyses from collaborative projects. Finally, in Section 5.4 I discuss some of the advantages and pitfalls of flexible modelling.

5.1 A flexible inference framework for linear mixed models

A large range of genetic models for quantitative traits can be cast as

$$\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{\beta},\, \boldsymbol{K}_{\boldsymbol{\theta}}\right), \tag{5.1}$$

where $\boldsymbol{y}$ is the phenotype vector for N samples, $\boldsymbol{X}$ is the $N \times K$ design matrix of K covariates, $\boldsymbol{\beta}$ is the vector of their effects and $\boldsymbol{K}_{\boldsymbol{\theta}}$ is the covariance function of the Gaussian model, which specifies the covariances between observations as a function of the model parameters $\boldsymbol{\theta}$ (see also Section D.1). Given an arbitrary covariance function $\boldsymbol{K}_{\boldsymbol{\theta}}$, model parameters are typically estimated by maximising the restricted maximum likelihood (Eq (A.7))

$$\mathcal{L} = \text{const} - \frac{1}{2}\log\det \boldsymbol{K}_{\boldsymbol{\theta}} - \log\det \boldsymbol{A}_{\boldsymbol{\theta}} - \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{\theta}} \tag{5.2}$$

where

$$\boldsymbol{A}_{\boldsymbol{\theta}} = \boldsymbol{X}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X} \tag{5.3}$$

$$\boldsymbol{\beta}_{\boldsymbol{\theta}} = \boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}. \tag{5.4}$$

Importantly, parameter inference in LIMIX is conducted by providing computational efficiency if possible. This is achieved by

- exploiting the structure of $\boldsymbol{K}_{\boldsymbol{\theta}}$,
- caching expensive computations during gradient-based parameter inference.

In the following, I discuss some of the implementation details of this inference framework.

I wrote most of the source code underlying the inference framework, with contributions from Danilo Horta and Barbara Rakitsch. In particular, Danilo Horta implemented an automatised caching strategy (which I do not discuss here) and worked on the testing and the deployment of the software.

5.1.1 Basic classes for inference

Within the LIMIX framework, the log (restricted) marginal likelihood (LML) is optimised using a gradient-based method, which requires repeated evaluations of the LML and its gradients (see Eqs A.7-A.8). Basic building blocks of these computations are

implemented in different classes, which are responsible for computing and caching these operations. The three basic classes for inference are:

- the mean term, which stores $\boldsymbol{X}$, $\boldsymbol{y}$ and $\boldsymbol{\beta_\theta}$ and is responsible for computing and caching all operations between these quantities;
- the covariance term, which is concerned with all the computations specific to the covariance $\boldsymbol{K_\theta}$, including Cholesky decomposition, matrix inverse, eigenvalue decomposition, log determinant (logdet) and gradients of the logdet (i.e. the trace term in Eq A.8);
- the Gaussian process (GP, see Section D.1 for a brief description of Gaussian processes), which governs the mean and the covariance terms. The GP evaluates the LML and its gradients and is responsible for computing all the operations that would involve cross talk between the mean term and the covariance term.

Note that the evaluation of the LML and its gradients requires the computation of the inverse, the logdet and the gradients of the logdet of $\boldsymbol{A_\theta} = \boldsymbol{X}^\top \boldsymbol{K_\theta}^{-1} \boldsymbol{X}$ (see Eqs A.7-A.8). As all these methods are already defined within the covariance term, it is convenient to introduce a derived covariance class, which we denote with Areml, that takes care of all computations involving $\boldsymbol{A_\theta}$. Areml is defined in the initialisation of the GP from mean and covariance terms.

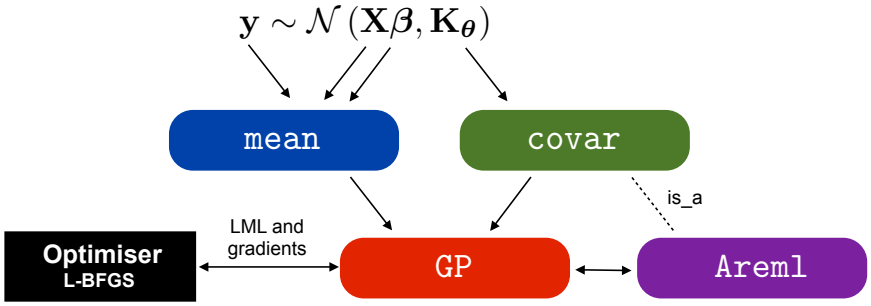


Figure 5.1: **Representation of the basic classes for inference in the LIMIX framework.** The figure illustrates the basic inference classes in the LIMIX framework and their dependencies. The mean term takes care of the computations between $\boldsymbol{y}$, $\boldsymbol{X}$ and $\boldsymbol{\beta_\theta}$, while the covariance term is responsible for all the computations concerning the covariance $\boldsymbol{K_\theta}$. The Areml term is a derived covariance class, which takes care of all the computations involving $\boldsymbol{A_\theta}$ in Eqs (A.7-A.8). Using the building blocks provided by the other classes, the GP allows for evaluating the LMM and its gradients and interfaces with the optimiser.

There are two main advantages of considering such a hierarchical structure for

inference. First, we can define derived classes of covariance terms and mean terms so that some of their methods can be adapted to specific cases, thereby enabling us to achieve faster computation (if possible, see Section 5.1.5). Second, we can define derived covariance classes that combine two or more covariance terms[1], for example, as discussed in Section 5.1.2, we define the sum and the Hadamart combinator class. Using this strategy, computations from low-level elements can be used by higher-level elements, resulting in a hierarchical structure that promotes reusability of the code and flexibility.

Gradient-based optimisation of the LML is performed considering the low-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) (Liu and Nocedal, 1989; Zhu et al., 1997), as implemented in the `fmin_l_bfgs_b` optimisation method in the SciPy python library (Jones et al., 2001). Specifically, the optimiser interfaces with the GP, which can evaluate the LML and its gradients. The GP communicates the updated parameter values to the mean term and the (composite) covariance term, which in turn propagates these updates to lower-level elements. After optimisation, standard errors of the model parameters can be also retrieved (see Section D.3 for details). A representation of the basic classes for inference and how they interact is given in **Fig. 5.1**.

5.1.2 Covariance models

A covariance matrix is defined as a real-valued positive semidefinite matrix. Specifically, a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is positive semidefinite if and only if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \geq 0$ $\forall \mathbf{x} \in \mathbb{R}^N$. The set of covariance matrices is closed under basic operations, such as matrix addition and Hadamart product. By exploiting this property, LIMIX provides the user with the possibility to combine basic covariance models to define more complex ones.

Basic covariance models. Basic covariance models include a fixed covariance (a predefined covariance matrix with a scaling parameter), a low-rank covariance, a freeform covariance (i.e. a general semi-definite positive matrix), a diagonal covariance, a squared exponential covariance and others. Details on the parametrisation of the different covariance terms implemented in LIMIX are provided in Section D.2. For a comprehensive review on covariance models, see Rasmussen (2006).

[1]While the same concept is applicable to mean terms, I do not discuss mean combinators here.

Combining covariance models. Let $\boldsymbol{K}_{1\boldsymbol{\theta}_1}$ and $\boldsymbol{K}_{2\boldsymbol{\theta}_2}$ denote two $N \times N$ covariance matrices with parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, respectively. The sum and the Hadamart product of the two covariance terms

$$\boldsymbol{K}_{\boldsymbol{\theta}}^{(\text{sum})} = \boldsymbol{K}_{1\boldsymbol{\theta}_1} + \boldsymbol{K}_{2\boldsymbol{\theta}_2} \tag{5.5}$$

$$\boldsymbol{K}_{\boldsymbol{\theta}}^{(\text{Had})} = \boldsymbol{K}_{1\boldsymbol{\theta}_1} \odot \boldsymbol{K}_{2\boldsymbol{\theta}_2}, \tag{5.6}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$, are valid covariance matrices. Additionally, the derivatives of $\boldsymbol{K}_{\boldsymbol{\theta}}^{(\text{sum})}$ and $\boldsymbol{K}_{\boldsymbol{\theta}}^{(\text{Had})}$ can be computed using chain rules

$$\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}^{(\text{sum})}}{\partial \boldsymbol{\theta}_j} = \begin{cases} \frac{\partial \boldsymbol{K}_{1\boldsymbol{\theta}_1}}{\partial \boldsymbol{\theta}_{1j}} & \text{if } j \leq |\boldsymbol{\theta}_1| \\ \frac{\partial \boldsymbol{K}_{2\boldsymbol{\theta}_2}}{\partial \boldsymbol{\theta}_{2(j-|\boldsymbol{\theta}_1|)}} & \text{otherwise} \end{cases} \tag{5.7}$$

$$\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}^{(\text{Had})}}{\partial \boldsymbol{\theta}_j} = \begin{cases} \frac{\boldsymbol{K}_{1\boldsymbol{\theta}_1}}{\partial \boldsymbol{\theta}_{1j}} \odot \boldsymbol{K}_{2\boldsymbol{\theta}_2} & \text{if } j \leq |\boldsymbol{\theta}_1| \\ \boldsymbol{K}_{1\boldsymbol{\theta}_1} \odot \frac{\boldsymbol{K}_{2\boldsymbol{\theta}_2}}{\partial \boldsymbol{\theta}_{2(j-|\boldsymbol{\theta}_1|)}} & \text{otherwise} \end{cases} \tag{5.8}$$

The LIMIX framework exploits these properties proposing a modelling framework where covariance models can be combined flexibly.

5.1.3 Kronecker-structured covariance models

Let us consider the covariance of the matrix-variate mixed model in Section 2.4

$$\underbrace{\boldsymbol{C}_g \otimes \boldsymbol{R}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{C}_n \otimes \boldsymbol{I}_N}_{\text{noise component}} . \tag{5.9}$$

Denoting with P the number of analysed traits and with N the number of individuals, $\boldsymbol{C}_g \in \mathbb{R}^{P \times P}$ and $\boldsymbol{C}_n \in \mathbb{R}^{P \times P}$ are the trait covariance matrices of the relatedness and the noise classes respectively, while $\boldsymbol{R}_g \in \mathbb{R}^{N \times N}$ denotes the realised relatedness matrix. Note that only $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ are estimated using restricted maximum likelihood, while $\boldsymbol{R}_g \in \mathbb{R}^{N \times N}$ is estimated from the genotype data. Within the LIMIX inference framework, the covariance in (5.9) is defined as a composite covariance term that combines parametrized covariance models $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$.

Let us now consider the covariance of the mtSet/iSet model in Eq (3.3)

$$\underbrace{\boldsymbol{C}_g \otimes \boldsymbol{G}\boldsymbol{G}^\top}_{\text{low-rank set component}} + \underbrace{\boldsymbol{C}_g \otimes \boldsymbol{R}_g}_{\text{relatedness component}} + \underbrace{\boldsymbol{C}_n \otimes \boldsymbol{I}_N}_{\text{noise component}} , \tag{5.10}$$

where an additional (low-rank) Kronecker product is considered. In Eq (5.10), we introduced the trait set covariance $\boldsymbol{C}_r \in \mathbb{R}^{P\times P}$ and the genotype matrix $\boldsymbol{G} \in \mathbb{R}^{N\times R}$ of the R variants in the considered set. Within LIMIX, the covariance in (5.10) is also defined as a composite covariance term that combines LIMIX covariance functions $\boldsymbol{C}_r$, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$. As many of computations are common to the two covariance models, the three-term covariance class reuses most methods implemented in two-term covariance class in the LIMIX framework.

One advantage of the framework is that the user can specify arbitrary parametrized covariance functions for $\boldsymbol{C}_r$, $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$. As as consequence, the models in iSet can be easily obtained using the covariance combinator in Eq (5.10) and setting (i) a freeform covariance for $\boldsymbol{C}_g$, (ii) a freeform covariance for $\boldsymbol{C}_n$, (iii) either a block, a rank-one or a freeform model for $\boldsymbol{C}_r$ (which correspond to the persistent, the rescaling-GxC and the general-GxC models in Section 4.1.1, respectively). Note that the framework also allows for more complex covariance models. For example, in analyses of time-series phenotypes, a squared exponential kernel (see Section D.2) could be used to model trait-to-trait covariances of the genetic component across time points.

Covariance matrices that are sum of multiple Kronecker terms can also be defined within LIMIX.

5.1.4 An example of complex covariance model to study social effects

In the previous sections, I discussed a framework that allows for combining multiple covariance functions. In this section, I showcase how this framework can help define complex LMMs for genetic analyses. In particular, I here discuss a statistical model I built in collaboration with Amelie Baud, a postdoc in Oliver Stegle's group, to study genetic effects from social partners (social genetic effect). Full details on the model and the genetic analysis can be found in Baud et al. (2017). Note that similar models have been considered for the study of social effects in previous studies (Bijma, 2014).

In social effect LMMs, the genetic contribution to the phenotype is modelled as the sum of a direct and a social component (Bijma, 2014). For each individual, the direct genetic component is defined as the additive effect of its own genes, while the social genetic component is defined as the additive effect of the genes of its social partners. As an illustrative example, let us consider an experimental design with $2N$ animals in N cages, with each cage containing two animals. Additionally, let $\boldsymbol{G} \in \mathbb{R}^{N\times S}$ denote the standardised genotype matrix for S variants and let us introduce the $N \times N$ cage

design matrix $\boldsymbol{Z}$ as

$$\boldsymbol{Z}_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are cagemates and } i \neq j \\ 0 & \text{otherwise} \end{cases} \tag{5.11}$$

The genetic contribution to the phenotype is defined as

$$\boldsymbol{g} = \underbrace{\boldsymbol{G}\boldsymbol{\beta}}_{\text{direct}} + \underbrace{\boldsymbol{Z}\boldsymbol{G}\boldsymbol{\gamma}}_{\text{social}}, \tag{5.12}$$

where

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0},\ \frac{1}{S}\boldsymbol{C}_g \otimes \boldsymbol{I}_S\right). \tag{5.13}$$

Here, $\boldsymbol{C}_g$ is a 2×2 covariance matrix modelling the covariance between the effects of the genes of one individual (focal individual) and the effects of the genes of its partner (on the focal individual). Marginalising out the random effects and introducing $\boldsymbol{R} = \frac{1}{S}\boldsymbol{G}\boldsymbol{G}^\top$ results in

$$\boldsymbol{g} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{K}^{(\text{ds})}(\boldsymbol{C}_g; \boldsymbol{R}, \boldsymbol{Z})\right), \tag{5.14}$$

where

$$\boldsymbol{K}^{(\text{ds})}(\boldsymbol{C}_g; \boldsymbol{R}, \boldsymbol{Z}) = \underbrace{\boldsymbol{C}_{g11}\boldsymbol{R}}_{\text{direct}} + \underbrace{\boldsymbol{C}_{g22}\boldsymbol{Z}\boldsymbol{R}\boldsymbol{Z}^T}_{\text{social}} + \underbrace{\boldsymbol{C}_{g12}\left(\boldsymbol{R}\boldsymbol{Z}^T + \boldsymbol{Z}\boldsymbol{R}\right)}_{\text{covariance}}. \tag{5.15}$$

Within the LIMIX framework, this covariance model can be defined from an arbitrary covariance function $\boldsymbol{C}_g$ and constant matrices $\boldsymbol{R}$ and $\boldsymbol{Z}$.

Following the same rationale, we introduced an environmental component with direct and indirect environmental effects[2] (see also Bijma, 2014). The total covariance of the model is

$$\underbrace{\boldsymbol{K}^{(\text{ds})}(\boldsymbol{C}_g; \boldsymbol{R}, \boldsymbol{Z})}_{\text{genetic component}} + \underbrace{\boldsymbol{K}^{(\text{ds})}(\boldsymbol{C}_n; \boldsymbol{I}, \boldsymbol{Z})}_{\text{environmental component}}, \tag{5.16}$$

where $\boldsymbol{C}_n$ is a 2×2 covariance matrix that models the covariance between direct and indirect environmental effects. Importantly, the implementation of such a complex covariance model is straightforward within the LIMIX framework, as it only requires the definition of the derived covariance model in Eq (5.15), which builds on an existing (and arbitrary) LIMIX covariance term ($\boldsymbol{C}_g$ in Eq (5.15)).

[2]Direct environmental effects correspond to iid noise. Indirect environmental effects correspond to effects that are mediated by the social partner.

5.1.5 Exploiting covariance structures

Naive inference in Gaussian models is bound to the $O(N^3)$ complexity of some of the computations to evaluate the LML and its gradients. These computations include the product of the inverse of $\boldsymbol{K}_{\boldsymbol{\theta}}$ by a matrix, the product of the gradients of $\boldsymbol{K}_{\boldsymbol{\theta}}$ by a matrix, the logdet of $\boldsymbol{K}_{\boldsymbol{\theta}}$ and the gradients of the logdet. Importantly, as we have seen in the previous chapters, some of these operations can be more efficiently computed for specific structures of the total covariance $\boldsymbol{K}_{\boldsymbol{\theta}}$. For example, if $\boldsymbol{K}_{\boldsymbol{\theta}}$ is the sum of a low-rank matrix of rank R and a diagonal matrix, its determinant can be computed in $O(NR^2 + R^3)$ using the matrix determinant lemma (Harville, 1998). Similarly, the operation $\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{W}$, where $\boldsymbol{W}$ is an $N \times M$ matrix, can be computed in $O(NKR + NR^2 + KR^2 + R^3)$ using the Woodbury identity (Woodbury, 1950).

Within this inference framework, all the methods whose computational complexity depends on the structure of the covariance matrix are defined within the covariance term. As a consequence of this design consideration, efficient inference comes automatically when the total covariance allows (and provides) efficient implementations of these methods (i.e. the product of the inverse of $\boldsymbol{K}_{\boldsymbol{\theta}}$ by a matrix, the product of the gradients of $\boldsymbol{K}_{\boldsymbol{\theta}}$ by a matrix, the logdet of $\boldsymbol{K}_{\boldsymbol{\theta}}$ and the gradients of the logdet). This strategy has been used to implement the iSet model for stratification analysis (see Section 4.4.1), for which the sole implementation of a specialised covariance term and its efficient methods (see Section A.6) enables inference with linear complexity in the number of individuals, $O(N)$. For some special models we have achieved further speedups by implementing specialised GP classes. This approach has been used for implementations of mtSet, mtSet-PC and mtSet-LowRankBg.

5.1.6 Other flexible inference frameworks

The idea of building composite covariance terms based on simpler building blocks is not exclusive to LIMIX. For example, GPy (GPy, since 2012) and PyGP (`https://github.com/PMBio/pygp`) also propose a flexible framework based on GPs, where covariance functions can be combined to define complex GP regression models. However, none of these implementations focuses on analysis needs in genetics. Specifically, these frameworks do not provide efficient implementations for Kronecker-structured covariance matrices, do not enable computation of standard errors and omit fixed effects.

In the context of genetics, softwares such as AsREML (Gilmour et al., 2009) and Wombat (Meyer, 2007) are widely popular in the animal breeding community. These software tools, especially AsREML, enable genetic analyses considering a large class of

linear mixed models. However, these methods are not open-source and do not implement recent mixed-models implementations to achieve fast computation in analyses of multiple variants and/or traits (Listgarten et al., 2013; Lippert et al., 2014a; Zhou and Stephens, 2014; Casale et al., 2015) (see also the comparison in Zhou and Stephens, 2014, Table 1).

5.2 Modules for genetic analyses

In the previous section, I discussed the basic framework for parameter inference. Here, I describe higher-level modules that implement recurrent analyses in genetic studies. Specifically, I discuss modules for variance decomposition and association testing across multiple traits. The material presented in this section is joint work with Christoph Lippert. I have developed all the code required for the variance decomposition module. Christoph implemented the code for fixed effect testing.

5.2.1 The variance decomposition module

Denoting with N the number of samples and with P the number of phenotypes, the variance component model implemented in LIMIX can be cast as

$$\boldsymbol{Y} = \sum_{j=1}^{J} \boldsymbol{F}_j \boldsymbol{B}_j \boldsymbol{A}_j + \sum_{i=1}^{I} \boldsymbol{U}_i + \boldsymbol{\Psi}, \tag{5.17}$$

where the $N \times P$ phenotype matrix $\boldsymbol{Y}$ is modelled as the sum of J fixed effect terms, I random effect terms and residual noise.

Fixed effects. Each of the fixed effect terms has a trait design matrix $\boldsymbol{A}_j \in \mathbb{R}^{M_j \times P}$, an individual design matrix $\boldsymbol{F}_j \in \mathbb{R}^{N \times K_j}$ and a fixed effect matrix $\boldsymbol{B}_j \in \mathbb{R}^{K_j \times M_j}$. Here, K_j is the number of covariates for term j and M_j is the number of independent effects that each of the K_j covariates has across the analysed traits. While the trait and the individual design matrices are known (i.e. they are specified by the user), the fixed effect matrices are estimated by restricted maximum likelihood. Importantly, the trait design $\boldsymbol{A}_j$ specifies the cross-trait architecture of the M_j independent effects. For example, the scenario in which each covariate in term j has the same effect on all the traits corresponds to $\boldsymbol{A}_j = \boldsymbol{1}_{1\times P}$ ($M_j = 1$). Alternatively, the scenario where each covariate has a different effect on each trait corresponds to $\boldsymbol{A}_j = \boldsymbol{I}_P$ ($M_j = P$). A more complex example in which the covariates have the same effects on all traits with

the exception of trait p corresponds to

$$\boldsymbol{A}_j = \begin{bmatrix} 1 & \cdots & 1 & \underbrace{0}_{p\text{-th column}} & 1 & \cdots & 1 \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix} \in \mathbb{R}^{2\times P}. \tag{5.18}$$

Random effects. Each of the random effects $\boldsymbol{U}_i$ follows a matrix-variate normal distribution with row covariance matrix $\boldsymbol{R}_i \in \mathbb{R}^{N\times N}$ and column covariance matrix $\boldsymbol{C}_i \in \mathbb{R}^{P\times P}$

$$\boldsymbol{U}_i \sim \mathrm{MVN}\left(\boldsymbol{0},\, \boldsymbol{R}_i,\, \boldsymbol{C}_i\right). \tag{5.19}$$

While the row covariance matrices are known [3], the trait covariance matrices are estimated by restricted maximum likelihood. Notably, user can specify the covariance model of the column covariance matrices for the different random effects. Finally, the residual noise is also matrix-variate normally distributed and covaries across traits while it is independent across individuals $\boldsymbol{\Psi} \sim \mathrm{MVN}\left(\boldsymbol{0},\, \boldsymbol{I}_N,\, \boldsymbol{C}_n\right)$.

Parameter inference. The model in Eq (5.17) can be equivalently written as

$$\mathrm{vec}\left(\boldsymbol{Y}\right) \sim \mathcal{N}\left(\sum_{j=1}^{J}\left(\boldsymbol{A}_j^\top \otimes \boldsymbol{F}_j\right)\mathrm{vec}\left(\boldsymbol{B}\right)_j,\, \sum_{i=1}^{I}\left(\boldsymbol{C}_i \otimes \boldsymbol{R}_i\right) + \boldsymbol{C}_n \otimes \boldsymbol{I}_N\right), \tag{5.20}$$

which is a special case of Eq (5.1). Parameter inference is performed using the framework described in Section 5.1.

Predictions The model can be used to predict contributions from individual terms (either fixed or random) for test individuals (out-of-sample individuals). Denoting with $N^\star$ the number of out-of-sample individuals and with $\boldsymbol{F}_j^\star \in \mathbb{R}^{N^\star\times K_j}$ the design matrix for the $N^\star$ individuals, the predictions from fixed effect j are

$$\boldsymbol{Y}_j^\star = \boldsymbol{F}_j^\star \hat{\boldsymbol{B}}_j \boldsymbol{A}_j, \tag{5.21}$$

where $\hat{\boldsymbol{B}}_j$ is the restricted maximum likelihood estimator (MLE) of $\boldsymbol{B}_j$. Additionally, denoting with $\boldsymbol{R}_i^\star \in \mathbb{R}^{N^\star\times N}$ the cross-covariance between the original N individuals and

[3] Row covariance matrices are normalised so that $\frac{1}{NP-1}\mathrm{tr}\left(\boldsymbol{P}\boldsymbol{R}_i\right) = 1,\ i = 1,\ldots,I$ so that the diagonal elements of the corresponding covariance can be interpreted as variance explained (Kostem and Eskin, 2013).

the $N^\star$ out-of-sample individuals (see also Section 2.3.6) the predictions from random effect i are

$$\boldsymbol{Y}_i^\star = \boldsymbol{R}_i^\star \text{vec}^{-1}\left(\left(\sum_{k=1}^{I} \hat{\boldsymbol{C}}_k \otimes \boldsymbol{R}_k + \hat{\boldsymbol{C}}_n \otimes \boldsymbol{I}_N\right)^{-1} \text{vec}\left(\boldsymbol{Y} - \sum_{j=1}^{J} \boldsymbol{F}_j^\star \hat{\boldsymbol{B}}_j \boldsymbol{A}_j\right)\right)\hat{\boldsymbol{C}}_i, \tag{5.22}$$

where $\hat{\boldsymbol{C}}_i$ and $\hat{\boldsymbol{C}}_n$ are the restricted MLE of $\boldsymbol{C}_i$ and $\boldsymbol{C}_n$, respectively.

5.2.2 Flexible fixed effect tests in multi-trait mixed models

The multivariate model used for single-variant testing has a fixed effect term for covariates, a fixed effect for the genetic variant being tested, a random effect to account for confounding and a noise term

$$\boldsymbol{Y} = \boldsymbol{F}^{(\text{cov})}\boldsymbol{B}^{(\text{cov})}\boldsymbol{A}^{(\text{cov})} + \boldsymbol{g}\boldsymbol{b}^\top\boldsymbol{A} + \boldsymbol{U} + \boldsymbol{\Psi}, \quad \text{with} \tag{5.23}$$

$$\boldsymbol{U} \sim \text{MVN}\left(\boldsymbol{0},\, \boldsymbol{R},\, \boldsymbol{C}_g\right), \qquad \boldsymbol{\Psi} \sim \text{MVN}\left(\boldsymbol{0},\, \boldsymbol{I},\, \boldsymbol{C}_n\right).$$

Here, $\boldsymbol{F}^{(\text{cov})} \in \mathbb{R}^{N\times K}$, $\boldsymbol{A}^{(\text{cov})} \in \mathbb{R}^{M^{(\text{cov})}\times P}$ and $\boldsymbol{B}^{(\text{cov})} \in \mathbb{R}^{K\times M^{(\text{cov})}}$ denote respectively the individual design matrix, the trait design matrix and the effect size matrix for K covariates, respectively. Additionally, $\boldsymbol{g}$ denote the genotype vector of the variant being tested, $\boldsymbol{A} \in \mathbb{R}^{M\times P}$ the trait design matrix of the genotype (M is the number of the independent effects on the traits, see also previous section), $\boldsymbol{b} \in \mathbb{R}^M$ the effect size vector of the M effects, $\boldsymbol{R}$ the realised relatedness matrix, $\boldsymbol{C}_g$ the trait polygenic covariance and $\boldsymbol{C}_n$ the trait noise covariance. Following Korte et al. (2012) and Furlotte and Eskin (2015), the trait matrices $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ are estimated under the no-association model ($\boldsymbol{b} = \boldsymbol{0}$) using the variance decomposition module (employing the efficient inference scheme described in Section A.2). By default, $\boldsymbol{A}^{(\text{cov})} = \boldsymbol{I}_P$ and $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$ are freeform covariance terms.

The single-variant test is perfomed by comparing the alternative hypothesis $\mathcal{H}_1 : \boldsymbol{A} = \boldsymbol{A}_1$ versus the null hypothesis $\mathcal{H}_0 : \boldsymbol{A} = \boldsymbol{A}_0$ using a log-likelihood ratio test. Importanly, LIMIX enables the user to specify both $\boldsymbol{A}_1$ and $\boldsymbol{A}_0$, thereby offering more flexibility than previous implementations (Korte et al., 2012; Zhou and Stephens, 2014; Furlotte and Eskin, 2015). The standard multivariate tests proposed in Korte et al. (2012) (see Section 2.4.3) can be obtained as special cases by choosing $\boldsymbol{A}_0$ and $\boldsymbol{A}_1$ appropriately (**Table 5.1**). However, LIMIX allows assessing more specific hypothesis on the trait design of the variant effect, as I show in the application case in the next

section.

Test	Alternative hypothesis	Null hypothesis
Any effect test	$\boldsymbol{A} = \boldsymbol{I}_P$	$\boldsymbol{A} = \boldsymbol{0}$
Common effect test	$\boldsymbol{A} = \boldsymbol{1}_P^\top$	$\boldsymbol{A} = \boldsymbol{0}$
Specific effect test for trait p	$\boldsymbol{A} = \begin{bmatrix} \mathcal{I}_p^\top \\ \mathcal{I}_{\sim p}^\top \end{bmatrix}$	$\boldsymbol{A}_0 = \boldsymbol{1}_P^\top$

Table 5.1: **Standard multivariate association tests and corresponding trait design in the null and in the alternative hypothesis.** Multivariate tests proposed in Korte et al. (2012) and corresponding choices of the trait design model in the null and in the alternative model. Here, $\mathcal{I}_d$ and $\mathcal{I}_{\sim d}$ are P-dimensional indicator vectors such that $(\mathcal{I}_p)_i = \delta_{ip}$ and $\mathcal{I}_{\sim p} = \boldsymbol{1}_D - \mathcal{I}_p$.

5.3 Vignettes

In this section, I describe two analyses from collaborative projects as vignettes of application, mainly focusing on the analysis challenges.

5.3.1 A genetic study of transcription initiation in Drosophila

This first application case concerns a genetic study of transcription initiation, profiled using Cap Analysis of Gene Expression (CAGE) in *Drosophila melanogaster* at three different stages of development. This analysis is part of a project in collaboration with Eileen Furlong's group in EMBL Heidelberg (Germany) and was published in Schor et al. (2017). CAGE measurements were available across three distinct stages of Drosophila development. The aim of this part of the analysis was to map QTLs for transcription initiation in a joint analysis across developmental stages.

The analysis of these data is challenging, as transcription initiation is a high-dimensional molecular trait consisting of hundreds of univariate measurements. Moreover, development adds another phenotypic dimension to the data. To approach these challenges, we first performed a conventional dimensionality reduction using PCA and then considered PC-based phenotypes for joint QTL mapping across developmental stages. The flexibility of the fixed-effect testing module implemented in LIMIX (see Section 5.2.2) was essential to define statistical tests for the specific study design (see below). The statistical pipeline for QTL mapping presented herein, was designed and validated by me in collaboration with Jacob Degner.

Background. A transcription start site (TSS) is a genomic location where transcription is initiated (Zvelebil and Baum, 2007). CAGE is a molecular profiling assay that enables the characterisation of transcription initiation on a genome-wide scale (Shiraki et al., 2003). CAGE isolates the sequence fragment at the 5' end of RNA molecules, which is the first part of the gene being transcribed. Mapping these short sequences to a reference genome enables the characterisation of the TSS distribution. Recent studies have shown that while many genes have a unique and well-defined TSS, for others, the distribution of TSS can span regions of up to thousands of bases (Lenhard et al., 2012; Carninci et al., 2006; Ni et al., 2010). While genetic effects on RNA expression levels have been largely studied, the extent to which genetic variation affects transcription initiation remains unknown. To investigate this, Eileen Furlong's group profiled transcription initiation in 81 genotyped lines of *Drosophila melanogaster* at 2-4h, 6-8h and 10-12h after egg laying using CAGE.

Phenotype definition. Transcription initiation regions (TIR) were defined as the 1kb regions centred around the highest CAGE peaks. In total, 13,508 TIR were identified. Denoting with N (= 81) the number of individuals and with D (= 3) the number of development stages, each TIR corresponds to $N \times D \times 1,000$ count measures (where 1,000 is the number of base pairs in a TIR). Even considering a single-stage analysis, joint modelling of the count data in a TRI would entail a joint QTL mapping of 1,000 univariate traits. To reduce the dimensionality of the problem, we projected the TSS distribution onto the three leading principal components, i.e. we performed dimensionality reduction in the base pair space. Specifically, we performed PCA on square-rooted counts across all lines and developmental stages. For each TIR, we also defined a mean-based phenotype as the sum of the read counts in the TIR. To adjust for batch effects and other hidden covariates, we applied PEER (Stegle et al., 2012) independently for each TIR and developmental stage to each of the $K_{\mathrm{PC}} = 3$ PC-based phenotypes and the mean-based phenotype considering 10 unknown factors. The residuals from PEER were quantile-normalised to a normal distribution to ensure that model assumptions were fulfilled.

Molecular QTL mapping. For each TIR, we considered all bi-allelic variants with MAF $> 5\%$ that are within 100 kb from the centres of TIR regions and considered three different analyses:

1. **Single-stage analysis of mean expression.** For each TIR and developmental stage, we considered the univariate linear mixed model described in Section 2.29.

We used the RRM to model genetic relatedness between lines.

2. **Multi-stage analysis of mean expression.** For each TIR, we considered the multi-trait linear mixed model in Eq (5.23) jointly modelling mean expression levels across the three developmental stages. We considered both a common effect test across all developmental stages and a specific effect test for each stage (see Section 5.2.2).

3. **Multi-stage analysis of PC-based phenotypes.** For each TIR, we considered the multi-trait linear mixed model in Eq (5.23) jointly modelling the $K_{\mathrm{PC}} = 3$ PC-based phenotypes across the $D = 3$ developmental stages, resulting in a total of 9 phenotypes. Denoting with $\boldsymbol{Y}_i \in \mathbb{R}^{N \times D}$ the matrix of PC i phenotypes, where rows are samples and columns are developmental stages, the total phenotype matrix is $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, \boldsymbol{Y}_3] \in \mathbb{R}^{N \times KD}$. Tests for common and specific effects across stages require the increased flexibility made available by the LIMIX framework. In this setting, a common effect is an effect that is heterogeneous across the K_{PC} PCs but constant (for each PC) across the D developmental stages. In the same vein, a specific effect at stage d is an effect that is different at stage d (for each PC) with respect to the other developmental stages. These two tests correspond to

$$\begin{aligned} \text{common effect test} \quad &: \quad \mathcal{H}_1 : \boldsymbol{A} = \boldsymbol{I}_K \otimes \mathbf{1}_{D\times 1} \quad \text{vs} \quad \mathcal{H}_0 : \boldsymbol{A} = \mathbf{0} \\ \text{specific effect test} \quad &: \quad \mathcal{H}_1 : \boldsymbol{A} = \boldsymbol{I}_K \otimes \begin{bmatrix} \mathcal{I}_d^\top \\ \mathcal{I}_{\sim d}^\top \end{bmatrix} \quad \text{vs} \quad \mathcal{H}_0 : \boldsymbol{A} = \boldsymbol{I}_K \otimes \mathbf{1}_{D,1} \end{aligned}$$

where $\mathcal{I}_d$ and $\mathcal{I}_{\sim d}$ are D-dimensional indicator vectors such that $(\mathcal{I}_d)_i = \delta_{id}$ and $\mathcal{I}_{\sim d} = \mathbf{1}_D - \mathcal{I}_d$.

To account for multiple testing we used a two-step procedure (see discussion in Section 2.2.2). First, we calculated a TIR-level P value for each TIR considering 10,000 permutations of the genotype data across lines. Second, the TIR-level P values were corrected for multiple testing across TIRs using the Benjamini-Hochberg (BH) procedure. For the single-stage analysis, the BH correction was performed both across TIRs and developmental stages.

Results. Joint modelling of mean-based phenotypes across multiple stages increased power compared to the single-stage analysis (**Fig. 5.2A-B**). Additionally, the multi-stage PC-based analysis almost doubled the number of significant TIRs with respect to

the mean-based analysis, identifying 4,526 TIRs with a significant QTL (FDR $< 1\%$, **Fig. 5.2A-B**). Note that this set of QTLs includes variants that only affect the expression level (see **Fig. 5.2C**), variants that only affects the shape of the TSS distribution (see **Fig. 5.2D**) and variants that affect both. Interestingly, the model did not retrieve any TIR with significant stage-specific effects.

5.3.2 Dissecting the genetic and the epigenetic component of gene expression

The analysis described in this section is a component of the BluePrint Whole Package 10 (BP-WP10, Chen et al. (2016)). BP-WP10 generated high-resolution transcriptional, genetic and epigenetic profiles in three immune cell types. These data offer a unique opportunity to study the interplay of genetic and epigenetic factors in the regulation of gene expression. The aim of the analysis presented herein is to assess the extent to which associations between epigenetic changes and gene expression are driven by underlying local genetic variation. The robust identification of these associations requires correction for different levels of sample heterogeneity, including sample processing, polygenic effects and underlying local genetic variation. To do so, we designed different LMMs with multiple variance components using LIMIX. The analysis presented here was done in collaboration with Nicole Soranzo's group at the Sanger Institute. I have designed the set of statistical models and analysed the data.

Data The BluePrint Whole Package 10 (BP-WP10) project has generated high-resolution transcriptional, genetic and epigenetic profiles in three immune cell types from 200 healthy Europeans, including $CD14^+$ monocytes, $CD16^+$ neutrophils and naive $CD4^+$ T cells. For brevity, I here consider only $CD14^+$ monocytes. The dataset consists of the following molecular layers:

- whole genome sequencing data: 5,237,919 common variants at MAF $> 4\%$;
- M-values for 440,905 CpG sites profiled with the 450K array;
- ChIP data for H3K27ac/H3K4me1 histone marks[4]: 64,843/39,815 peaks;
- RNA-seq: 16,577 genes.

[4]H3K4me1 indicates the mono-methylation of the K4 lysine of histone H3 while H3K27ac indicates the acetylation of the K4 lysine of histone H3. H3K4me1 has been associated to regions with active and poised enhancers while H3K27ac is specific to active enhancers (Hon et al., 2009; Creyghton et al., 2010).

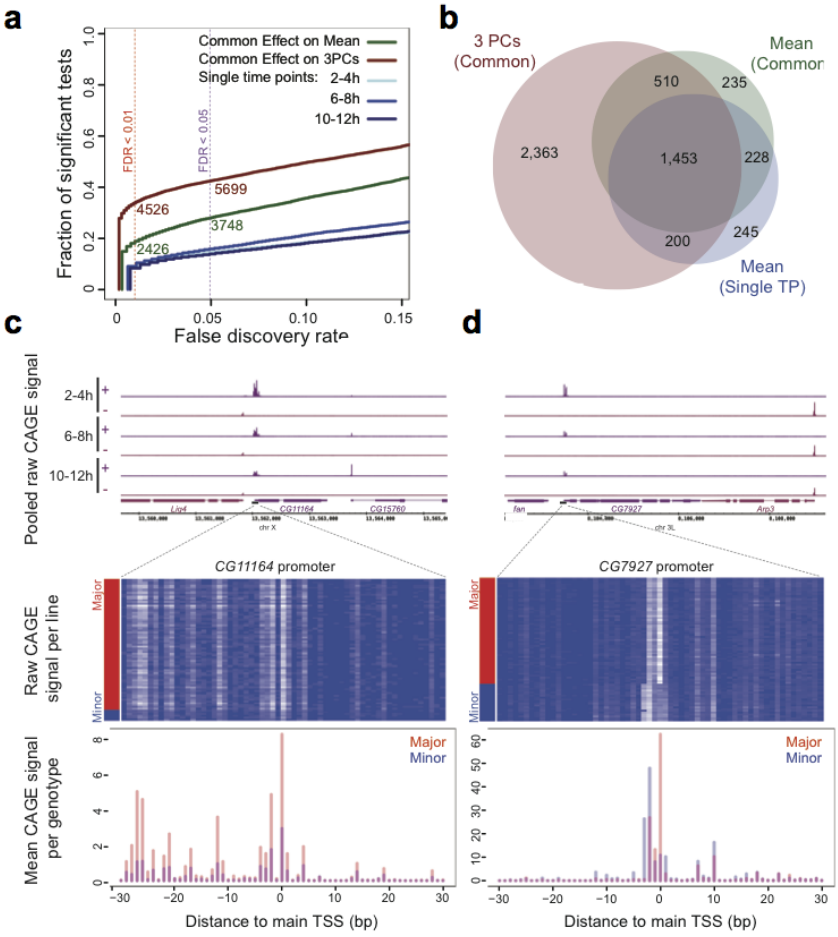


Figure 5.2: **Results from the QTL mapping of transcription initiation in Drosophila using multi-trait mixed models.**
Figure by Jacob Degner and Ignacio Shor. Used with permission.
(**a**) Fraction of significant TIRs with a QTL considering alternative association tests as a function of the FDR cut-off. Association tests include single-stage association analysis for mean expression, common effect test on mean expression and common effect test on PC-based phenotypes. (**b**) Overlap of genes having significant QTL (FDR<1%) considering the three association tests. (**c**) Example of QTL that affects the expression level without affecting the shape of the TSS distribution. The top panel shows the raw CAGE signal for the different developmental stages pulled across all lines. The mid panel shows a heatmap of the raw CAGE signal at 6-8h for each line, where individuals are ordered based on the genotype at the lead variant. The third panel shows the mean CAGE signal stratifying by the genotype of the lead QTL. (**d**) Analogous representation as in (**c**) for a QTL that affects the shape of the distribution of transcription initiation.

To account for hidden covariates and confounding factors, we applied PEER (Stegle et al., 2012) to both the expression and the epigenetic data. In the analyses described in the following, we considered the set of 158 individuals for which all molecular layers had passed all quality control steps. Details on the different molecular assays and the preprocessing of the data can be found in Section D.4.1.

Mapping of *cis* epiQTL. To identify epigenetic associations with gene expression that are explainable in terms of local genetic effects, we tested for association between gene expression and epigenetic features within 1Mb from the gene-body either with or without accounting for local genetic variation.

- ***cis* epiQTL mapping.**

  For each gene, we considered the following LMM

$$\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{1}\mu + \boldsymbol{e}\beta, \quad \underbrace{\sigma_g^2 \boldsymbol{K}_g}_{\text{relatedness component}} + \underbrace{\sigma_h^2 \boldsymbol{K}_h}_{\text{expr. heterogeneity}} + \underbrace{\sigma_e^2 \boldsymbol{I}}_{\text{noise}}\right). \tag{5.24}$$

  Denoting with N the number of individuals, here $\boldsymbol{y} \in \mathbb{R}^N$ is the normalised[5] gene-expression vector, $\mathbf{1}\mu \in \mathbb{R}^N$ an intercept term, $\boldsymbol{e} \in \mathbb{R}^N$ the epigenotype vector of the feature being tested and $\boldsymbol{K}_g \in \mathbb{R}^{N\times N}$ the RRM. Additionally, $\boldsymbol{K}_h \in \mathbb{R}^{N\times N}$ denotes the expression heterogeneity (EH) covariance and was introduced to account for confounding due to hidden sample heterogeneity. Specifically, the EH covariance is defined as $\boldsymbol{K}_h = \boldsymbol{Z}\boldsymbol{Z}^T \in \mathbb{R}^{N\times N}$, where $\boldsymbol{Z}$ denotes the $N \times G$ gene expression matrix for N individuals and all G genes. The underlying idea of this strategy is that genome-wide expression heterogeneity is likely to capture technical confounding. This is a common strategy in QTL mapping (Kang et al., 2008a; Fusi et al., 2012). Importantly, we find that a variance component model not accounting for this contribution gave inflated variance component estimates for epigenetic features (Section D.4.2). When testing for association, we used a conservative approach and quantile-normalised epigenetic features to a normal distribution. For computational efficiency, the association testing was performed by fixing the relative contribution of the relatedness and the EH component to the total phenotypic variance. Specifically, we first used the null model (Eq (5.24) with $\beta = 0$) to estimate σ_g^2 and σ_h^2 by restricted maximum likelihood. Then we considered the univariate LMM in Section 2.29 and we set the covariance matrix

[5] quantile-normalised to a unit variance normal distribution

of the random effect to

$$\frac{\hat{\sigma}_g^2 \boldsymbol{K}_g + \hat{\sigma}_h^2 \boldsymbol{K}_h}{\hat{\sigma}_g + \hat{\sigma}_h}, \tag{5.25}$$

where $\hat{\sigma}_g^2$ and $\hat{\sigma}_h^2$ are the restricted MLE of σ_g^2 and σ_h^2 respectively under the null model.

- **Accounting for cis-genetic effects.** To account for local genetic effects, we first corrected epigenetic features as described in the following. For each epigenetic mark, we considered the linear mixed model

$$\boldsymbol{e} \sim \mathcal{N}\left(\mathbf{1}\mu + \boldsymbol{e}\beta,\ \sigma^2_{\text{g100kb}} \boldsymbol{R}_{\text{g100kb}} + \sigma_e^2 \boldsymbol{I}\right), \tag{5.26}$$

where $\boldsymbol{R}_{\text{g100kb}}$ denotes a local RRM built from genetic variants within 100kb from the epigenetic mark. The effect from local genetic variants was estimated using the best linear unbiased predictor (see Section 2.3.6) and the residuals were used as an estimate of the non-genetic component of the epigenetic marks (cisG-corrected features).

For each gene, we then considered the following LMM

$$\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{1}\mu + \boldsymbol{e}^\star\beta,\ \underbrace{\sigma^2_{\text{geno}} \boldsymbol{K}_{\text{geno}}}_{\textit{cis}\text{ genetic comp.}} + \underbrace{\sigma_g^2 \boldsymbol{K}_g}_{\text{relatedness component}} + \underbrace{\sigma_h^2 \boldsymbol{K}_h}_{\text{expr. heterogeneity}} + \underbrace{\sigma_e^2 \boldsymbol{I}}_{\text{noise}}\right), \tag{5.27}$$

where $\boldsymbol{e}^\star$ denotes the cisG-corrected epigenetic vector and $\boldsymbol{K}_{\text{geno}}$ denotes the local RRM built considering all variants in 1Mb from the gene-body. This additional random effect accounts for cis genetic variation within the entire 1Mb region considered in the mapping. We used the same strategy as above for the association testing. CisG-corrected epigenetic features were quantile-normalised to a Gaussian distribution prior to the association testing.

For multiple hypothesis correction, we performed a two-step procedure (as in Battle et al., 2014, see also discussion in Section 2.2.2). We first obtained a gene-level P-value as the minimum nominal P-value (Bonferroni-corrected to account for multiple testing across cis features) and then used the BH procedure to correct for multiple testing across genes. We called genes with significant epigenetic association at FDR $< 5\%$.

Results Accounting for cis-genetic variation reduced the number of genes with a significant epigenetic association from 5,813 to 2,861 (see **Fig. 5.3a-c**). The biological

interpretation of the results is ongoing work and will be omitted for brevity.

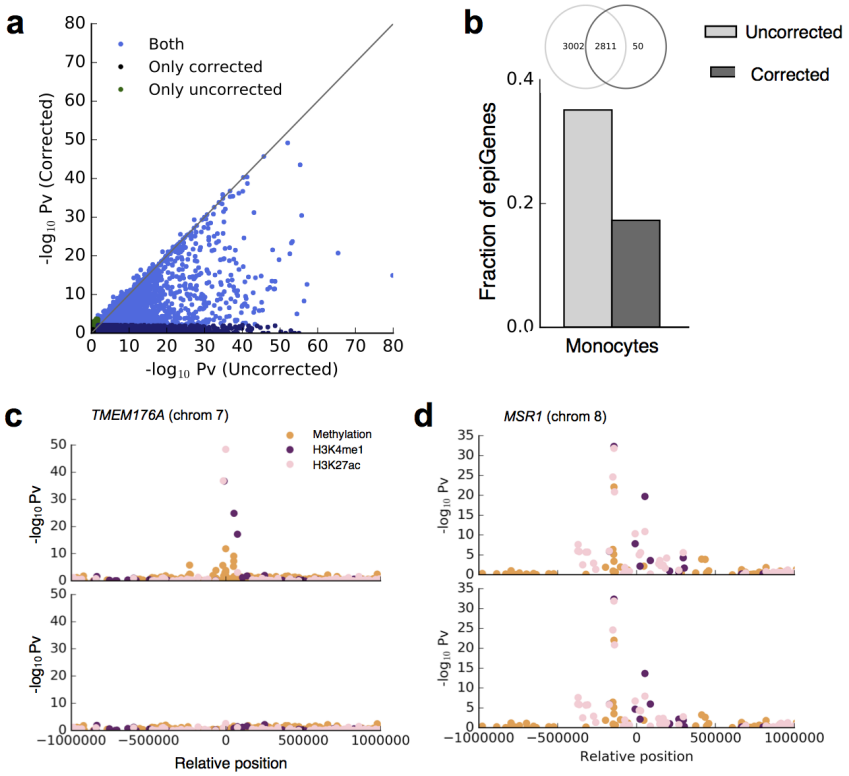


Figure 5.3: **Results from the analysis in the BluePrint data.** (**a**) Scatter plot of the gene-level P values obtained from the epiQTL mapping either accounting (y-axis) or not (x-axis) for cis-genetic variation. (**b**) Fraction of genes with significant epigenetic associations (FDR<5%) before and after accounting for cis-genetic variation. (**c**) Manhattan plot for gene *TMEM176A* obtained from the EWAS analysis on gene expression without (top panel) and with (bottom panel) accounting for cis-genetic variation. (**d**) Analogous plot as in (**c**) for gene *MSR1*.

5.4 Summary and discussion

The generation of cohorts with increasingly deep phenotype data posits the need for analysis tools that allow for adapting genetic models to variable study designs. Indeed, the analysis of such datasets typically requires iterative changes to the model, starting from a simple univariate approach, followed by more complex multivariate models. In this chapter, I presented LIMIX, a flexible mixed-model framework to address such analysis needs. Notably, LIMIX unifies a wide range of tasks in a single framework, including variance decomposition, single-variant and set-based association testing, and genomic predictions for the analysis of single and multiple traits. Technically, LIMIX builds on a general inference framework for Gaussian models, in which model parameters are estimated by maximising the restricted marginal likelihood. A key component of this inference framework is the covariance term, as it defines the set of model parameters and is a major determinant of the computational complexity of the model. Additionally, existing covariance models can be combined using basic operations; a strategy that facilitates implementation of new models and that promotes flexibility and re-usability of the code.

In addition to this inference framework, LIMIX provides higher-level modules that implement standard genetic analyses. I have described a module for flexible variance decomposition and a module for multi-trait GWAS that is more general than previous implementations. I have illustrated the utility of these modules in two applied projects. In the first application case, I described a genetic analysis of transcription initiation in *Drosophila melanogaster* during development. Using the multi-trait GWAS module in LIMIX, we defined a set of non-standard multivariate tests that consider common and development-specific genetic effects on transcription initiation. The flexibility of the LIMIX GWAS module was essential to adapt the statistical tests to the specific design. In the second application case, I discussed an integrative analysis of genetic, epigenetic and expression profiles. This analysis required the fitting of multiple LMMs accounting for different sources of confounding, which were implemented in a coherent fashion using the variance decomposition module in LIMIX.

Although flexible tools in genetics can help the analysis of complex datasets, flexibility comes at the cost of the increased complexity of the software. Indeed, in order for the user to leverage the flexibility of the framework and make optimal modelling choices, she or he will be required to have a good understanding of the modelling framework and some of its implementation details, a compromise that might be suboptimal for many users. One possible solution is to utilise a flexible framework as the foundation

for different predefined workflows. This strategy would enable (i) the development of new methods (by extending the core inference framework and by designing new high-level modules), (ii) the flexible design of genetic models (by using higher-level modules such as the variance decomposition) and (iii) the use of predefined workflows.

6 | Concluding remarks

The success of the first genome-wide association studies, which uncovered common variants with moderate effect sizes for traits such as type 2 diabetes (Scott et al., 2007) and coronary artery disease (Burton et al., 2007), fuelled the hope that GWAS could be applied to characterise the genetic component of virtually any trait of interest (Visscher et al., 2012). However, as increasing numbers of traits have been surveyed using genetic association analyses, it has become increasingly clear that these expectations were too optimistic and that the genotype-to-phenotype map is more complex than initially hypothesised. Many traits of interest are complex and can be regulated by tens or even hundreds of genetic loci with each locus having a small individual effect (Wood et al., 2014; Ripke et al., 2014). Moreover, genetic variants commonly entail pleiotropic effects across several related phenotypes and diseases (Fortune et al., 2015; Pickrell et al., 2016) and interactions of genetic effects with environmental factors and other contexts are common (Andreassi, 2009; Winkler et al., 2015). Additionally, many discovered variants lie in non-coding regions of the genome, hampering the interpretation of the molecular mechanisms that underlie such associations. Genetic studies of gene expression levels and other molecular traits have helped identify the regulated gene for a fraction of such intergenic GWAS loci. However, as the effects of genetic variants can arise in specific tissue types or under specific stimuli (GTEx Consortium, 2015; Fairfax et al., 2014), these genetic studies need to be conducted in disease-relevant cellular states.

One avenue to tackle these challenges is to increase statistical power by considering ever-increasing sample sizes (Visscher et al., 2012). This strategy is employed in large meta- and mega-analyses, where several studies exceed sample sizes of 100,000 individuals (Gormley et al., 2016; Sudlow et al., 2015). A complementary direction of investigation is to consider joint analyses across multiple traits, molecular layers and contexts. These integrative analyses can help characterise pleiotropy (Fortune et al., 2015), assess colocalisation of genetic effects on gene expression and complex

traits (Wallace et al., 2012) and study context-specificity (Flutre et al., 2013). As increasingly deep phenotype data are being generated, there is the need for new integrative methods for tying together different traits and genetic effects in flexible ways, while retaining computational efficiency in large cohorts.

This thesis has contributed new statistical methods for integrative analysis. In Chapter 3, I presented mtSet, a mixed-model approach that enables association testing of variant-sets with multiple traits. This approach can leverage the availability of multiple related phenotypes in the same individuals, while modelling effects from sets of variants at the same locus and accounting for confounding. In applications to a number of experimental settings, we find that joint modelling of multiple traits and variants offers power advantages compared to methods that aggregate either across traits or variants in isolation. Importantly, mtSet uses an efficient algorithm implementation, allowing for applications to large cohorts. Building on this inference scheme, in Chapter 4, I derived a new strategy to test for interactions between a set of genetic variants and categorical contexts (iSet). iSet accounts for polygenic effects and allows for characterising context-specificity at specific loci. In an application to a monocyte eQTL dataset, the proposed approach is better powered than a single-variant interaction test, suggesting that gene expression is often regulated by multiple causal variants. Moreover, our results reveal that changes of configuration of causal variants between contexts are common. I also extended the iSet method to enable interaction testing in GWAS cohorts when using context variables to stratify individuals into distinct subgroups. Finally, in Chapter 5, I presented LIMIX, a mixed-model software suite that enables different types of genetic analyses. LIMIX provides the user with the flexibility to design customised genetic models, facilitating integrative analyses in specific data designs. To illustrate the use of this modelling framework, I have presented two vignettes. In the first vignette, I showcased the use of flexible fixed effect testing to investigate context-specific genetic effects in a joint analysis of multiple contexts and traits. The second use case illustrates flexible variance component modelling, which I used to account for different types of confounding in association testing. In addition to methodological contributions, the application of the proposed methods to different data has also yielded new insights into the genetic architecture of traits. Perhaps most notably, these results highlight that even low-level molecular traits, such as gene expression, have surprisingly complex *cis* genetic architectures, with multiple associated variants and complex changes between different cellular contexts.

There are a number of use cases for the methods developed in this thesis that could be explored in the future. The proposed association set test offers the potential for

the discovery of new quantitative trait loci using set-based analyses of up to tens of related phenotypes. Examples of interesting applications include joint genetic analyses of multiple metabolite measurements, groups of phenotypes that are predictive for disease susceptibility, measures of stress, anxiety and other behavioural traits in animal models, and genetic molecular analyses of small gene networks or correlated epigenetic factors. Importantly, as mtSet can account for different types of genetic relatedness, it can be applied to either large human cohorts of unrelated individuals, such as UK Biobank (Sudlow et al., 2015) and NFBC (Sabatti et al., 2009), or medium-sized cohorts of related individuals, such as the human cohorts considered in Sidore et al. (2015) and Panoutsopoulou et al. (2014) and the model organism datasets considered in Baud et al. (2014) and Atwell et al. (2010).

The proposed interaction set test will enable region-based gene-context (GxC) interaction analyses in the two most common designs of GxC studies: complete designs, where every individual has been phenotyped in all contexts, and stratified designs, where each phenotype has been measured in only one of a number of contexts. Examples of analyses with a complete design include GxC studies of gene expression and other molecular traits across cell types, tissues or development, analyses of global phenotypes in model organisms across different environments or stimuli (e.g., see designs in Sasaki et al. (2015) and Bloom et al. (2013)), and genetic analyses of traits over time in longitudinal studies. Conversely, the possibility to consider stratified designs will enable GxC interaction analyses in large human cohorts by stratifying individuals by sex, age bins and other context variables.

The flexible LIMIX software tool will enable the design of ad hoc genetic models to address specific analysis needs. Applications of LIMIX have concerned variance component analysis of multiple molecular layers in human primary cells (Chen et al., 2016), gene-environment (GxE) interaction study of methylation and flowering time in Arabidopsis (Dubin et al., 2015b; Sasaki et al., 2015), genetic analysis of transcription initiation (Schor et al., 2017), multivariate expression locus mapping in Drosophila during development (Cannavò et al., 2016), studies of social genetic effects (Baud et al., 2017) and predictions of growth traits in yeast (Märtens et al., 2016). The availability of robust software implementations of these methods will enable new genetic analyses in ambitious study designs to gain insights into the complexities of the genotype-phenotype map.

I will conclude with an outlook on future research by discussing some extensions and follow-up analyses linked to the work in this thesis. First, while I here have focused on sliding-window experiments, different strategies to define variant-sets can be adopted

to test for specific hypotheses and increase statistical power. For example, sets can be defined by using genome annotations such as gene positions or ENCODE elements. The most common analysis types that implements this principle are gene-based set tests. This strategy offers the advantage that the obtained results can be directly interpreted and increases power by reducing the number of tested hypothesis and leveraging LD. Analyses of variant-sets could also be applied to study the local genetic architecture at specific loci, for example, by testing for residual effects from multiple low-frequency alleles after conditioning on the main genetic signals. Such analyses could be easily implemented within mtSet by modelling the main effect as fixed and considering a set relatedness matrix built from only low-frequency variants. Additionally, it would be important to explore the optimal design of variant-sets by comparing alternative ways of weighting the variants in the set component, for example, to account for uneven LD tagging (Speed et al., 2012), to prioritise low-frequency variants (Wu et al., 2011) or to introduce prior biological knowledge (MacLeod et al., 2016).

As efficient multi-trait implementations are currently limited to analyses of tens of traits, one important line of research is the development of new models for the genetic analysis of high-dimensional phenotypes, such as high throughput molecular measurements and images of cells, faces and organs. Such challenge can be tackled using different analysis strategies. For example, one approach is to design dimensionality reduction methods that can extract heritable latent variables by leveraging genotype information. Alternatively, one can extend current multi-trait implementations so that they can be directly applied to analyse hundreds of traits, which entails different statistical challenges. First, it will be important to devise analysis strategies to deal with missing phenotype observations. Promising avenues are approximate inference methods such as variational Bayes (Dahl et al., 2016) and efficient inference schemes available for Gaussian models with structured covariances (Wilson et al., 2014). Another statistical challenge is the robust estimation of large trait-to-trait covariance matrices. Inference of large covariances is an active area of research in statistics and principles of low-rank approximations, regularisation (Dahl et al., 2013) and bootstrap-based estimation could also be applied.

An important extension of the models presented in this thesis, which require access to genotype and phenotype population datasets, is to enable the analysis of summary statistic data from previous GWAS and molecular studies. Methods based on summary statistics are appealing as these data are more easily accessible, without the privacy concerns that limit access to genetic data (Pasaniuc and Price, 2016). For example, a probabilistic model on the variant effect statistics can be obtained from a generative

model on phenotype by using the relationship between trait observations and the estimated effect sizes (see also Chen et al., 2015). This approach could be used to extend the iSet model to consider applications on summary statistics from GTeX, enabling analyses of eQTL signal heterogeneity across multiple tissues.

While I here have focused on the study of genetic interactions with simple environments, specifically categorical contexts, the development of methods to study interactions with continuous and multidimensional environments is an important area of future work. For example, principles of variable warping (Fusi et al., 2014b) can be used to learn non-linear functions of environmental factors and determine on which scale continuous environments interact with genetic variants. Another interesting direction is to apply variance component models to study the heterogeneity of genetic signals across multiple correlated environmental exposures. These advanced GxE models will enable us to leverage the rich information on lifestyle covariates and other environments available in large cohorts such as UK Biobank, with the potential for revealing new insights on how environment mediates the genetic architecture of traits.

Finally, although several of the developed methods are available through open software (`https://github.com/limix/limix`) with accessible public interface documentation (`http://limix.readthedocs.io`) and interactive notebooks (`https://github.com/limix/limix-tutorials`), more work is needed to provide workflows for users to reuse and adapt. For example, in order to broaden the user community of the LIMIX software suite, it will be important to release command-line interface tools for an increasingly large set of genetic analyses.

A | Derivations

A.1 Restricted maximum likelihood

Denoting with the N the number of samples and with $\boldsymbol{y} \in \mathbb{R}^N$ the outcome vector, let us consider the Gaussian model

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{K}) \tag{A.1}$$

where $\boldsymbol{X}$ is the $N \times F$ design matrix of K fixed effects, $\boldsymbol{\beta}$ their effect sizes, $\boldsymbol{K}_{\boldsymbol{\theta}}$ the covariance function of the Gaussian model and $\boldsymbol{\theta}$ its parameters. The logarithm of the restricted marginal likelihood of the Gaussian model can be expressed as (Harville, 1974; LaMotte, 2007)

$$\begin{aligned} \mathcal{L} &= -\frac{N-F}{2}\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{K}_{\boldsymbol{\theta}} - \log\det \boldsymbol{A}_{\boldsymbol{\theta}} && \text{(A.2)} \\ &\quad -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{\theta}})^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{\theta}}) && \text{(A.3)} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{A}_{\boldsymbol{\theta}} &= \boldsymbol{X}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \boldsymbol{X} && \text{(A.4)} \\ \boldsymbol{\beta}_{\boldsymbol{\theta}} &= \boldsymbol{A}_{\boldsymbol{\theta}}^{-1} \boldsymbol{X}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}. && \text{(A.5)} \end{aligned}$$

An equivalent form that we will use in the following is

$$\begin{aligned} \mathcal{L} &= -\frac{N-F}{2}\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{K}_{\boldsymbol{\theta}} - \log\det \boldsymbol{A}_{\boldsymbol{\theta}} && \text{(A.6)} \\ &\quad -\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{\theta}} && \text{(A.7)} \end{aligned}$$

Gradients

The gradient of the restricted marginal likelihood with respect to $\boldsymbol{\theta}_i$ is

$$\begin{aligned}\frac{\partial \mathcal{L}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i} = & - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{A}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\right) + \frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y} \\ & - \boldsymbol{y}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{\theta}} - \frac{1}{2}\boldsymbol{\beta}_{\boldsymbol{\theta}}^\top \frac{\partial \boldsymbol{A}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\boldsymbol{\beta}_{\boldsymbol{\theta}} \end{aligned} \tag{A.8}$$

where

$$\frac{\partial \boldsymbol{A}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i} = -\boldsymbol{X}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X} \tag{A.9}$$

A.2 Implementation of LMMs with two-Konecker covariance matrices

Let us consider the model

$$\boldsymbol{Y} = \underbrace{\sum_{j=1}^{J} \boldsymbol{F}_j \boldsymbol{B}_j \boldsymbol{A}_j}_{\text{fixed effects}} + \underbrace{\boldsymbol{U}}_{\text{confounding}} + \underbrace{\boldsymbol{\Psi}}_{\text{noise}}, \tag{A.10}$$

where the $N \times P$ phenotype matrix $\boldsymbol{Y}$ is modelled as the sum of J fixed effect terms, a random effect to correct for confounding and residual noise. Each of the fixed effect terms has a trait design matrix $\boldsymbol{A}_j \in \mathbb{R}^{M_j \times P}$, an individual design matrix $\boldsymbol{F}_j \in \mathbb{R}^{N \times K_j}$ and a fixed effect matrix $\boldsymbol{B}_j \in \mathbb{R}^{K_j \times M_j}$. The terms $\boldsymbol{U} \in \mathbb{R}^{N \times P}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{N \times P}$ follow matrix-variate normal distributions

$$\boldsymbol{U} \sim \mathrm{MVN}\left(\boldsymbol{0},\, \boldsymbol{R}_g,\, \boldsymbol{C}_g\right), \qquad \boldsymbol{\Psi} \sim \mathrm{MVN}\left(\boldsymbol{0},\, \boldsymbol{I}_N,\, \boldsymbol{C}_n\right), \tag{A.11}$$

where $\boldsymbol{R}_g$ is a constant $N \times N$ covariance. Note that this model generalises the matrix-variate mixed model introduced in Section 2.4 and it is a particular case of the variance decomposition model considered in LIMIX (see Eq (5.17)). In restricted maximum likelihood the model parameters are defined by the covariance and, in this case, are the parameters of $\boldsymbol{C}_g$ and $\boldsymbol{C}_n$. In the following, to simplify notation, we do not explicitly indicate the dependency on the model parameters.

Using the properties of the vec operator and Kronecker product introduced in Sec-

tion 2.4.1 the model can be written as

$$\operatorname{vec}(\boldsymbol{Y}) \sim \mathcal{N}\left(\sum_{j=1}^{J}\left(\boldsymbol{A}_j^\top \otimes \boldsymbol{F}_j\right)\operatorname{vec}(\boldsymbol{B})_j\,,\, \boldsymbol{C}_g \otimes \boldsymbol{R}_g + \boldsymbol{C}_n \otimes \boldsymbol{I}_N\right). \tag{A.12}$$

This model is a special case of Eq (A.1) with

$$\begin{aligned}
\boldsymbol{y} &= \operatorname{vec}(\boldsymbol{Y}) && \text{(A.13)}\\
\boldsymbol{X} &= \begin{bmatrix}\boldsymbol{A}_1^T \otimes \boldsymbol{F}_1 & \dots & \boldsymbol{A}_J^T \otimes \boldsymbol{F}_J\end{bmatrix} \in \mathbb{R}^{NP\times K} && \text{(A.14)}\\
\boldsymbol{\beta} &= \begin{bmatrix}\operatorname{vec}(\boldsymbol{B}_1)\\ \vdots \\ \operatorname{vec}(\boldsymbol{B}_J)\end{bmatrix} && \text{(A.15)}\\
\boldsymbol{K} &= \boldsymbol{C}\otimes\boldsymbol{R} + \boldsymbol{\Sigma}\otimes\boldsymbol{I}, && \text{(A.16)}
\end{aligned}$$

where $K = \sum_j K_j M_j$.

In Section 2.4 we derived that the inverse of the covariance can be rewritten as (see Eq (2.82))

$$\begin{aligned}
\boldsymbol{K}^{-1} &= \boldsymbol{L}^T\boldsymbol{D}\boldsymbol{L} && \text{(A.17)}\\
\boldsymbol{L} &= \boldsymbol{L}_c \otimes \boldsymbol{L}_r. && \text{(A.18)}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{C}_g^\star &= \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2}\boldsymbol{U}_{\boldsymbol{C}_n}^\top \boldsymbol{C}_g \boldsymbol{U}_{\boldsymbol{C}_n}\boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2} && \text{(A.19)}\\
\boldsymbol{L}_c &= \boldsymbol{U}_{\boldsymbol{C}_g^\star}^\top \boldsymbol{S}_{\boldsymbol{C}_n}^{-1/2}\boldsymbol{U}_{\boldsymbol{C}_n}^\top && \text{(A.20)}\\
\boldsymbol{L}_r &= \boldsymbol{U}_{\boldsymbol{R}_g}^\top && \text{(A.21)}\\
\boldsymbol{L} &= \boldsymbol{L}_c \otimes \boldsymbol{L}_r && \text{(A.22)}\\
\boldsymbol{D} &= \left(\boldsymbol{S}_{\boldsymbol{C}_g^\star}\otimes\boldsymbol{S}_{\boldsymbol{R}_g} + \boldsymbol{I}_{NP}\right)^{-1}. && \text{(A.23)}
\end{aligned}$$

In the following, we will see how the restricted log marginal likelihood and its gradients can be computed efficiently exploiting the particular structure of the covariance.

Restricted log marginal Likelihood

Using Eq (A.17), the restricted log marginal likelihood in (A.7) can be written as

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\theta}} &= -\frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\log|\boldsymbol{A}| - \frac{1}{2}\text{vec}\left(\tilde{\boldsymbol{Y}}\right)^{\top}\boldsymbol{D}^{-1}\text{vec}\left(\tilde{\boldsymbol{Y}}\right) && \text{(A.24)}\\ &\quad + \frac{1}{2}\text{vec}\left(\tilde{\boldsymbol{Y}}\right)^{\top}\boldsymbol{D}^{-1}\tilde{\boldsymbol{X}}\boldsymbol{\beta}, && \text{(A.25)}\end{aligned}$$

where we have introduced

- The transformed phenotype vector

$$\text{vec}\left(\tilde{\boldsymbol{Y}}\right) = (\boldsymbol{L}_c \otimes \boldsymbol{L}_r)\,\text{vec}\,(\boldsymbol{Y}) = \text{vec}\left(\boldsymbol{L}_r\boldsymbol{Y}\boldsymbol{L}_c^T\right), \tag{A.26}$$

  whose computation has complexity $O(N^2P + NP^2)$;

- the transformed fixed effect design

$$\tilde{\boldsymbol{X}} = (\boldsymbol{L}_c \otimes \boldsymbol{L}_r)\,\boldsymbol{X} = \left[\tilde{\boldsymbol{A}}_1^T \otimes \tilde{\boldsymbol{F}}_1 \quad \ldots \quad \tilde{\boldsymbol{A}}_J^T \otimes \tilde{\boldsymbol{F}}_J\right]. \tag{A.27}$$

  Here we have introduced $\tilde{\boldsymbol{A}}_j = \boldsymbol{A}_j\boldsymbol{L}_c^{\top}$ and $\tilde{\boldsymbol{F}}_j = \boldsymbol{L}_r\boldsymbol{F}_j$. Computation of $\tilde{\boldsymbol{A}}_j$ has complexity $O(P^2M_j)$ and computation of $\tilde{\boldsymbol{F}}_j$ has complexity $O(N^2K_j)$.

Note that $\boldsymbol{L}_r$ (the row rotation matrix, see Section 2.4.4) is constant, i.e. does not change during parameter optimisation. As consequence all row rotations need to be performed only once prior to model fitting. These operations are highlighted in blue.

Eqs (A.4-A.5) can be written as

$$\begin{aligned}\boldsymbol{A} &= \tilde{\boldsymbol{X}}^{\top}\boldsymbol{D}^{-1}\tilde{\boldsymbol{X}} && \text{(A.28)}\\ \boldsymbol{\beta} &= \boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^{\top}\boldsymbol{D}^{-1}\text{vec}\left(\tilde{\boldsymbol{Y}}\right) = \boldsymbol{A}^{-1}\begin{bmatrix}\text{vec}(\tilde{\boldsymbol{F}}_1^{\top}\boldsymbol{D}^{-1}\tilde{\boldsymbol{Y}}\,\tilde{\boldsymbol{A}}_1^T)\\ \vdots\\ \text{vec}(\tilde{\boldsymbol{F}}_J^{\top}\boldsymbol{D}^{-1}\tilde{\boldsymbol{Y}}\,\tilde{\boldsymbol{A}}_J^T)\end{bmatrix}. && \text{(A.29)}\end{aligned}$$

$\boldsymbol{A}$ can be computed in $O(NPK + NPK^2)$ and $\boldsymbol{\beta}$ in $O(\sum_j K_jNP + \sum_j K_jPM_j)$.

All terms in Eq (A.25) can be derived with complexity $O(K^3 + NPK)$.

Gradients

Note that the term $\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}$ can be written as

$$\begin{aligned}\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1} &= \boldsymbol{L}\boldsymbol{D}\underbrace{\left(\boldsymbol{L}^\top\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}\right)}_{\tilde{\boldsymbol{C}}\otimes\tilde{\boldsymbol{S}}}\boldsymbol{D}\boldsymbol{L}^\top & \text{(A.30)}\\ &= \boldsymbol{L}\boldsymbol{D}\left(\tilde{\boldsymbol{C}}\otimes\tilde{\boldsymbol{S}}\right)\boldsymbol{D}\boldsymbol{L}^\top, & \text{(A.31)}\end{aligned}$$

where we used that as both $\boldsymbol{L}$ and $\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}$ are Kronecker products, then $\boldsymbol{L}^\top\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}$ will also be a Kronecker product that we denoted with $\tilde{\boldsymbol{C}}\otimes\tilde{\boldsymbol{S}}$. Note that $\tilde{\boldsymbol{S}}$ is either $\boldsymbol{S}_{\boldsymbol{R}_g}$ or $\boldsymbol{I}$ depending on whether $\boldsymbol{\theta}_i$ is a parameter of $\boldsymbol{C}_g$ or $\boldsymbol{C}_n$ respectively.

All terms in (A.8) can be computed efficiently as follows.

- The first term is

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) &= \operatorname{tr}\left(\boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\top\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) & \text{(A.32)}\\ &= \operatorname{tr}\left(\boldsymbol{D}^{-1}\boldsymbol{L}^\top\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}\right) & \text{(A.33)}\\ &= \operatorname{diag}(\boldsymbol{D}^{-1})^\top\operatorname{diag}\left(\tilde{\boldsymbol{C}}\otimes\tilde{\boldsymbol{S}}\right) & \text{(A.34)}\\ &= \operatorname{diag}(\boldsymbol{D}^{-1})^\top\left(\operatorname{diag}\left(\tilde{\boldsymbol{C}}\right)\otimes\operatorname{diag}\left(\tilde{\boldsymbol{S}}\right)\right), & \text{(A.35)}\end{aligned}$$

  which has complexity $O(P^3 + NP)$.

- The second term is

$$\operatorname{tr}\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\theta}_i}\right) = \operatorname{tr}\left(\boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^\top\boldsymbol{D}^{-1}\left(\tilde{\boldsymbol{C}}\otimes\tilde{\boldsymbol{S}}\right)\boldsymbol{D}^{-1}\tilde{\boldsymbol{X}}\right), \quad \text{(A.36)}$$

  which has complexity $O(NPK + P^2NK + NPK^2 + K^3)$.

- The third term is

$$\boldsymbol{y}^\top\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\boldsymbol{y} = \underbrace{\operatorname{vec}\left(\tilde{\boldsymbol{Y}}\right)^\top\boldsymbol{D}^{-1}}_{\text{already computed}}\left(\tilde{\boldsymbol{C}}\otimes\tilde{\boldsymbol{S}}\right)\underbrace{\boldsymbol{D}^{-1}\operatorname{vec}\left(\tilde{\boldsymbol{Y}}\right)}_{\text{already computed}}, \quad \text{(A.37)}$$

  which has complexity $O(NP + NP^2)$.

- The fourth term is

$$\boldsymbol{y}^\top \boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}^{-1} \boldsymbol{X}\boldsymbol{\beta} = \underbrace{\text{vec}\left(\tilde{\boldsymbol{Y}}\right)^\top \boldsymbol{D}^{-1}}_{\text{already computed}} \left(\tilde{\boldsymbol{C}} \otimes \tilde{\boldsymbol{S}}\right) \underbrace{\boldsymbol{D}^{-1}\tilde{\boldsymbol{X}}\boldsymbol{\beta}}_{\text{already computed}}, \tag{A.38}$$

  which has complexity $O(NP + NP^2)$.

- The fifth term is

$$\boldsymbol{\beta}^\top \frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\theta}_i} \boldsymbol{\beta} = \underbrace{\boldsymbol{\beta}^\top \tilde{\boldsymbol{X}}^\top \boldsymbol{D}^{-1}}_{\text{already computed}} \underbrace{\left(\tilde{\boldsymbol{C}} \otimes \tilde{\boldsymbol{S}}\right) \boldsymbol{D}^{-1} \tilde{\boldsymbol{X}}\boldsymbol{\beta}}_{\text{already computed}}, \tag{A.39}$$

  which has complexity $O(NP)$.

A.3 Implementation of mtSet gradients

The efficient implementation of the marginal likelihood are given in Section 3.1.3. Here we discuss the efficient calculation of the gradients. The derivative of the log likelihood with respect to the column covariance parameter $\theta_i \in \boldsymbol{\theta}$ is given by

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_i} = -\frac{1}{2}\text{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) + \frac{1}{2}\text{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\text{vec}\left(\boldsymbol{Y}\right), \tag{A.40}$$

where

$$\boldsymbol{K} \;=\; \boldsymbol{E}\boldsymbol{E}^\top \otimes \boldsymbol{G}\boldsymbol{G}^\top + \boldsymbol{C}_g \otimes \boldsymbol{R}_g + \boldsymbol{C}_n \otimes \boldsymbol{I}_N. \tag{A.41}$$

Here, $\boldsymbol{G} \in \mathbb{R}^{N\times R}$ denotes the standardised genotype matrix for R variants in the set, $\boldsymbol{E} \in \mathbb{R}^{P\times C}$ is defined such that $\boldsymbol{E}\boldsymbol{E}^\top$ is the trait covariance of the set component ($C < P$), $\boldsymbol{C}_g \in \mathbb{R}^{P\times P}$ is the trait covariance of the relatedness component, $\boldsymbol{C}_n \in \mathbb{R}^{P\times P}$ is the trait covariance of the noise component and $\boldsymbol{R}_g \in \mathbb{R}^{N\times N}$ is the RRM (see also Section 3.1.3). In the following, we will use the result in Eq (3.26):

$$\boldsymbol{K}^{-1} = \boldsymbol{L}^\top \boldsymbol{D}\boldsymbol{L} - \boldsymbol{L}^\top \boldsymbol{D}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^\top \boldsymbol{D}\boldsymbol{L}$$

to derive the efficient computation of the gradients.

Evaluating the trace term The first term can be rewritten as

$$\begin{aligned}
\mathrm{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) &= \mathrm{tr}\left(\left[\boldsymbol{L}^T\left(\boldsymbol{D}-\boldsymbol{D}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^T\boldsymbol{D}\right)\boldsymbol{L}\right]\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{D}-\boldsymbol{D}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^T\boldsymbol{D}\right)\boldsymbol{L}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}^\top\right) \\
&= \mathrm{tr}\left(\left(\boldsymbol{D}-\boldsymbol{D}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^T\boldsymbol{D}\right)\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\right) \\
&= \mathrm{tr}\left(\boldsymbol{D}\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\right)-\mathrm{tr}\left(\boldsymbol{D}\boldsymbol{W}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^T\boldsymbol{D}\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\right) \\
&= \sum_{jk}\left(\boldsymbol{D}\odot\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\right)_{jk}-\sum_{jk}\left(\boldsymbol{\Lambda}^{-1}\odot\boldsymbol{W}^T\boldsymbol{D}\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\boldsymbol{D}\boldsymbol{W}\right)_{jk} \\
&= \mathrm{diag}(\boldsymbol{D})^T\mathrm{diag}(\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}})-\sum_{jk}\left(\boldsymbol{\Lambda}^{-1}\odot\widehat{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\right)_{jk}
\end{aligned}$$

where

$$\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}} = \boldsymbol{L}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}^T \tag{A.42}$$

$$= \boldsymbol{L}_c\boldsymbol{C}_{\theta_i}\boldsymbol{L}_c^\top\otimes\boldsymbol{L}_r\boldsymbol{R}_i\boldsymbol{L}_r^\top. \tag{A.43}$$

$$= \begin{cases}\boldsymbol{L}_c\frac{\partial(\boldsymbol{E}\boldsymbol{E}^T)}{\partial\boldsymbol{\theta}_i}\boldsymbol{L}_c^T\otimes\boldsymbol{W}_r\boldsymbol{W}_r^T & \text{if } \theta_i \text{ is an entry of } \boldsymbol{E}\\ \boldsymbol{L}_c\frac{\partial(\boldsymbol{C_g})}{\partial\boldsymbol{\theta}_i}\boldsymbol{L}_c^T\otimes\boldsymbol{S}_r & \text{if } \theta_i \text{ is a param of } \boldsymbol{C}_g\\ \boldsymbol{L}_c\frac{\partial(\boldsymbol{C_n})}{\partial\boldsymbol{\theta}_i}\boldsymbol{L}_c^T\otimes\boldsymbol{I} & \text{if } \theta_i \text{ is a param of } \boldsymbol{C}_n\end{cases} \tag{A.44}$$

$$= \begin{cases}\widetilde{\frac{\partial\boldsymbol{E}}{\partial\boldsymbol{\theta}_i}}\widetilde{\boldsymbol{E}}^T+\widetilde{\boldsymbol{E}}\widetilde{\frac{\partial\boldsymbol{E}}{\partial\boldsymbol{\theta}_i}}^\top\otimes\boldsymbol{W}_r\boldsymbol{W}_r^T & \text{if } \theta_i \text{ is an entry of } \boldsymbol{E}\\ \widetilde{\frac{\partial(\boldsymbol{C_g})}{\partial\boldsymbol{\theta}_i}}\otimes\boldsymbol{S}_r & \text{if } \theta_i \text{ is a param of } \boldsymbol{C}_g\\ \widetilde{\frac{\partial(\boldsymbol{C_n})}{\partial\boldsymbol{\theta}_i}}\otimes\boldsymbol{I} & \text{if } \theta_i \text{ is a param of } \boldsymbol{C}_n\end{cases}, \tag{A.45}$$

where $\widetilde{\boldsymbol{E}} = \boldsymbol{L}_c\boldsymbol{E}$, $\widetilde{\frac{\partial\boldsymbol{E}}{\partial\boldsymbol{\theta}_i}} = \boldsymbol{L}_c\frac{\partial\boldsymbol{E}}{\partial\boldsymbol{\theta}_i}$ and

$$\tag{A.46}$$

$$\widehat{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}} = \boldsymbol{W}^T\boldsymbol{D}\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\boldsymbol{D}\boldsymbol{W}. \tag{A.47}$$

First, we compute the column covariance matrix of $\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}$, which can be done in $O(P^2C)$, $O(P^3)$ and $O(P^3)$ respectively for random effect, error and noise parameters. Calcu-

lating $\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}$ requires more care: we first compute the dot product between $\boldsymbol{D}$ and $\boldsymbol{W}$, which requires us to explicitly calculate $\boldsymbol{W}$, taking $O(NPRC)$ time and $O(NPRC)$ space. The resulting matrix consists of NP rows and RC columns.

In the next step, we multiply each column of $\boldsymbol{DW}_{:,\boldsymbol{i}}$ with $\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}$ from the right side exploiting the same tricks as in (3.34):

$$\begin{aligned}
\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}} \boldsymbol{DW}_{:,\boldsymbol{i}} &= \begin{cases} \left(\widetilde{\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\theta}_i}} \widetilde{\boldsymbol{E}}^T + \widetilde{\boldsymbol{E}} \widetilde{\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\theta}_i}}^\top \otimes \boldsymbol{W}_r \boldsymbol{W}_r^T \right) \boldsymbol{DW}_{:,\boldsymbol{i}} & \text{if } \theta_i \text{ is an entry of } \boldsymbol{E} \\ \left(\widetilde{\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i}} \otimes \boldsymbol{S}_r \right) \boldsymbol{DW}_{:,\boldsymbol{i}} & \text{if } \theta_i \text{ is a param of } \boldsymbol{C}_g \\ \left(\widetilde{\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i}} \otimes \boldsymbol{I} \right) \boldsymbol{DW}_{:,\boldsymbol{i}} & \text{if } \theta_i \text{ is a param of } \boldsymbol{C}_n \end{cases} \\
&= \begin{cases} \operatorname{vec}\left(\boldsymbol{W}_r \boldsymbol{W}_r^T \operatorname{vec}^{-1}\left(\boldsymbol{DW}_{:,\boldsymbol{i}}\right) \left(\widetilde{\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\theta}_i}} \widetilde{\boldsymbol{E}}^T + \widetilde{\boldsymbol{E}} \widetilde{\frac{\partial \boldsymbol{E}}{\partial \boldsymbol{\theta}_i}}^\top \right) \right) & \text{if } \theta_i \text{ is a parameter of } \boldsymbol{E} \\ \operatorname{vec}\left(\boldsymbol{S}_r \operatorname{vec}^{-1}\left(\boldsymbol{DW}_{:,\boldsymbol{i}}\right) \widetilde{\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i}} \right) & \text{if } \theta_i \text{ is a parameter of } \boldsymbol{C}_g \\ \operatorname{vec}\left(\operatorname{vec}^{-1}\left(\boldsymbol{DW}_{:,\boldsymbol{i}}\right) \widetilde{\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i}} \right) & \text{if } \theta_i \text{ is a parameter of } \boldsymbol{C}_n. \end{cases}
\end{aligned} \tag{A.48}$$

This leads to an overall runtime complexity of $O(RC \cdot (NPC + NCR))$, $O(RC \cdot (NP + NP^2))$ and $O(RCNP^2)$ for region, random effect and region parameters.

We use the same trick to compute the multiplication between $\boldsymbol{W}^T$ and $\boldsymbol{D}\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}\boldsymbol{DW}$ efficiently, leading to a complexity of $O(RC(NPC + NCR))$. Finally, computing the trace term has an additional runtime of $O(NP + C^2R^2)$.

Evaluating the derivative of the squared form The derivative of the squared form can be rewritten as

$$\begin{aligned}
&\operatorname{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}^{-1} \operatorname{vec}\left(\boldsymbol{Y}\right) \\
&\quad = \operatorname{vec}\left(\tilde{\boldsymbol{Y}}\right)^T \left(\boldsymbol{D} - \boldsymbol{DW}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^T\boldsymbol{D}\right) \widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}} \left(\boldsymbol{D} - \boldsymbol{DW}\boldsymbol{\Lambda}^{-1}\boldsymbol{W}^T\boldsymbol{D}\right) \operatorname{vec}\left(\tilde{\boldsymbol{Y}}\right) \\
&\quad = \left(\operatorname{vec}\left(\hat{\boldsymbol{Y}}\right) - \boldsymbol{DW}\boldsymbol{\Lambda}^{-1}\operatorname{vec}\left(\bar{\boldsymbol{Y}}\right)\right)^\top \widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}} \left(\operatorname{vec}\left(\hat{\boldsymbol{Y}}\right) - \boldsymbol{DW}\boldsymbol{\Lambda}^{-1}\operatorname{vec}\left(\bar{\boldsymbol{Y}}\right)\right).
\end{aligned}$$

where $\hat{\boldsymbol{Y}} = \boldsymbol{D}\operatorname{vec}\left(\tilde{\boldsymbol{Y}}\right)$. We start by multiplying $\boldsymbol{\Lambda}^{-1}$ with $\operatorname{vec}\left(\bar{\boldsymbol{Y}}\right)$, which can be done in $O(C^2R^2)$ after having precomputed the inverse. Exploiting that $\boldsymbol{W}$ has Kronecker structure and $\boldsymbol{D}$ is a diagonal matrix, reduces the runtime for multiplying the resulting matrix with $\boldsymbol{DW}$ from the left from $O(NP + NPRC)$ to $O(NPC + RNC + NP)$. In the next step, we subtract the resulting vector from $\operatorname{vec}\left(\hat{\boldsymbol{Y}}\right)$ and multiply it with $\widetilde{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}}$ from the left, having an additional runtime of $O(NPR + NP^2)$, $O(NP + NP^2)$ and $O(NP^2)$ for the region, the random effect and the noise term respectively. Finally

we have to multiply two vectors of size $N \times P$, which can be done in $O(NP)$ time. A tabular overview of the individual computations and how often these need to be carried out can be found in Table **??**.

A.4 Implementation of mtSet-PC

Denoting with with $\boldsymbol{Y}$ the $N \times P$ phenotype matrix for N samples and P traits, with $\boldsymbol{F} \in \mathbb{R}^{N \times K}$ the sample design matrix of the fixed effect, with $\boldsymbol{G} \in \mathbb{R}^{N \times R}$ the standardised genotype matrix for the R variants in the set, with $\boldsymbol{C}_r \in \mathbb{R}^{P \times P}$ the trait covariance matrix of the set component and with $\boldsymbol{C}_n$ the residual trait covariance matrix, the model considered by mtSet-PC is

$$\operatorname{vec}(\boldsymbol{Y}) \sim \mathcal{N}\left((\boldsymbol{I} \otimes \boldsymbol{F}) \operatorname{vec}(\boldsymbol{B}), \boldsymbol{C}_r \otimes \boldsymbol{G}\boldsymbol{G}^\top + \boldsymbol{C}_n \otimes \boldsymbol{I}_N \right), \tag{A.49}$$

where analogous to mtSet, we set $\boldsymbol{C}_r = \boldsymbol{E}\boldsymbol{E}^T$ where $\boldsymbol{E} \in \mathbb{R}^{P \times C}$ and $C < P$ (i.e. $\boldsymbol{C}_r$ is low-rank). In mtSet-PC, population structure is modelled as fixed effects using the first principal components instead than using a random effect term as in the full mtSet. Again, this is a special form of the model in Eq (A.1) with

$$\begin{aligned} \boldsymbol{y} &= \operatorname{vec}(\boldsymbol{Y}) && \text{(A.50)} \\ \boldsymbol{X} &= (\boldsymbol{I}_P \otimes \boldsymbol{F}) \in \mathbb{R}^{NP \times K} && \text{(A.51)} \\ \boldsymbol{K} &= \underbrace{\boldsymbol{C}_r \otimes \boldsymbol{G}\boldsymbol{G}^T}_{\text{low rank}} + \boldsymbol{C}_n \otimes \boldsymbol{I}_N, && \text{(A.52)} \end{aligned}$$

Let us consider the log restricted marginal likelihood in (A.3):

$$\begin{aligned} \mathcal{L} &= -\frac{N-F}{2}\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{K} - \log\det \boldsymbol{A} && \text{(A.53)} \\ & \quad -\frac{1}{2}(\operatorname{vec}(\boldsymbol{Y}) - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{K}^{-1} \underbrace{(\operatorname{vec}(\boldsymbol{Y}) - \boldsymbol{X}\boldsymbol{\beta})}_{\operatorname{vec}(\boldsymbol{Z})} && \text{(A.54)} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{X}^\top \boldsymbol{K}^{-1} \boldsymbol{X} && \text{(A.55)} \\ \boldsymbol{\beta} &= \boldsymbol{A}^{-1} \boldsymbol{X}^\top \boldsymbol{K}^{-1} \operatorname{vec}(\boldsymbol{Y}). && \text{(A.56)} \end{aligned}$$

The covariance matrix can be rewritten as

$$\begin{aligned}
\boldsymbol{K} &= \boldsymbol{E}\boldsymbol{E}^T \otimes \boldsymbol{G}\boldsymbol{G}^T + \boldsymbol{C}_n \otimes \boldsymbol{I}_N && \text{(A.57)}\\
&= \left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right)\left(\boldsymbol{E}^\star \boldsymbol{E}^{\star T} \otimes \boldsymbol{G}\boldsymbol{G}^T + \boldsymbol{I}_{NP}\right)\left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right)^T && \text{(A.58)}\\
&= \left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right) \times && \text{(A.59)}\\
&\quad \left(\left(\boldsymbol{U}_{\boldsymbol{E}^\star} \otimes \boldsymbol{U}_{\boldsymbol{G}}\right)\left(\boldsymbol{S}_{\boldsymbol{E}^\star} \otimes \boldsymbol{S}_{\boldsymbol{G}}\right)\left(\boldsymbol{U}_{\boldsymbol{E}^\star} \otimes \boldsymbol{U}_{\boldsymbol{G}}\right)^T + \boldsymbol{I}_{NP}\right) \times && \text{(A.60)}\\
&\quad \left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right)^T && \text{(A.61)}
\end{aligned}$$

where we used the notation $\boldsymbol{M} = \boldsymbol{U}_{\boldsymbol{M}} \boldsymbol{S}_{\boldsymbol{M}}^{1/2} \boldsymbol{X}_{\boldsymbol{M}}$[1] for the singular value decomposition of $\boldsymbol{M}$. The inverse of $\boldsymbol{K}$ can be written as

$$\begin{aligned}
\boldsymbol{K}^{-1} &= \left(\boldsymbol{S}_n^{-1/2} \boldsymbol{U}_n^T \otimes \boldsymbol{I}_N\right)^T \times && \text{(A.62)}\\
&\quad \left(\boldsymbol{I}_{NP} - \left(\boldsymbol{U}_{\boldsymbol{E}^\star} \otimes \boldsymbol{U}_{\boldsymbol{G}}\right) \underbrace{\left(\boldsymbol{S}_{\boldsymbol{E}^\star}^{-1} \otimes \boldsymbol{S}_{\boldsymbol{G}}^{-1} + \boldsymbol{I}_{RC}\right)^{-1}}_{\boldsymbol{D}} \underbrace{\left(\boldsymbol{U}_{\boldsymbol{E}^\star} \otimes \boldsymbol{U}_{\boldsymbol{G}}\right)^T}_{\boldsymbol{W}}\right) \times && \text{(A.63)}\\
&\quad \underbrace{\left(\boldsymbol{S}_n^{-1/2} \boldsymbol{U}_n^T \otimes \boldsymbol{I}_N\right)}_{\boldsymbol{L}}\\
&= \boldsymbol{L}^T \left(\boldsymbol{I} - \boldsymbol{W}^T \boldsymbol{D} \boldsymbol{W}\right) \boldsymbol{L} && \text{(A.64)}
\end{aligned}$$

Calculating the SVD of $\boldsymbol{E}^\star$ and $\boldsymbol{G}^\star$ takes respectively $O(PC^2)$ and $O(NR^2)$ operations. All computations that has to be performed only once during optimisation are marked in blue.

Evaluating the log-likelihood The log-likelihood of the model is

$$\mathcal{L}_{\boldsymbol{\theta}} = \text{const.} - \frac{1}{2} \underbrace{\text{vec}\left(\boldsymbol{Z}\right)^T \boldsymbol{K}^{-1} \text{vec}\left(\boldsymbol{Z}\right)}_{\text{squared form term}} - \frac{1}{2} \underbrace{\log\det \boldsymbol{K}}_{\text{logdet term}} - \frac{1}{2} \underbrace{\log\det \boldsymbol{A}}_{\text{reml term}} \quad \text{(A.65)}$$

The log-determinant term can be computed as follows by applying the matrix de-

[1] $\boldsymbol{M} \in \mathbb{R}^{n_1, n_2}$, $\boldsymbol{U} \in \mathbb{R}^{n_1, n_1}$, $\boldsymbol{S} \in \mathbb{R}^{n_1, n_2}$, $\boldsymbol{X} \in \mathbb{R}^{n_2, n_2}$

terminant lemma

$$\begin{aligned}
\log\det \boldsymbol{K} &= \log\det\left(\left(\boldsymbol{U}_{\boldsymbol{E}^\star}\otimes\boldsymbol{U}_{\boldsymbol{G}}\right)\left(\boldsymbol{S}_{\boldsymbol{E}^\star}\otimes\boldsymbol{S}_{\boldsymbol{G}}\right)\left(\boldsymbol{U}_{\boldsymbol{E}^\star}\otimes\boldsymbol{U}_{\boldsymbol{G}}\right)^T+\boldsymbol{I}_{NP}\right) && \text{(A.66)}\\
&\quad +N\log\det\boldsymbol{S}_n && \text{(A.67)}\\
&= \log\det\left(\boldsymbol{S}_{\boldsymbol{E}^\star}^{-1}\otimes\boldsymbol{S}_{\boldsymbol{G}}^{-1}+\boldsymbol{I}\right)+R\log\det\boldsymbol{S}_{\boldsymbol{E}^\star}+C\log\det\boldsymbol{S}_{\boldsymbol{G}} && \text{(A.68)}\\
&\quad +N\log\det\boldsymbol{S}_n. && \text{(A.69)}
\end{aligned}$$

$\boldsymbol{A}$ and $\boldsymbol{\beta}$ can be computed respectively as

$$\begin{aligned}
\boldsymbol{A} &= \boldsymbol{X}^T\boldsymbol{K}^{-1}\boldsymbol{X} && \text{(A.70)}\\
&= (\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{L}\boldsymbol{X}-(\boldsymbol{W}\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}(\boldsymbol{W}\boldsymbol{L}\boldsymbol{X}) && \text{(A.71)}\\
&= (\boldsymbol{L}_c\otimes\boldsymbol{F})^T(\boldsymbol{L}_c\otimes\boldsymbol{F})-(\boldsymbol{W}_c\boldsymbol{L}_c\otimes\boldsymbol{W}_r\boldsymbol{F})^T\boldsymbol{D}(\boldsymbol{W}_c\boldsymbol{L}_c\otimes\boldsymbol{W}_r\boldsymbol{F}) && \text{(A.72)}
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{\beta} &= \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right)= && \text{(A.73)}\\
&= \boldsymbol{A}^{-1}\left((\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right)-(\boldsymbol{W}\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right)\right) && \text{(A.74)}\\
&= \boldsymbol{A}^{-1}\left(\mathrm{vec}\left(\boldsymbol{F}^T\boldsymbol{Y}\boldsymbol{L}_c^T\boldsymbol{L}_c\right)-(\boldsymbol{W}\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}\mathrm{vec}\left(\boldsymbol{W}_r\boldsymbol{Y}\boldsymbol{L}_c^T\boldsymbol{W}_c^T\right)\right) && \text{(A.75)}
\end{aligned}$$

Finally, the quadratic term can be rewritten as

$$\begin{aligned}
\mathrm{vec}\left(\boldsymbol{Z}\right)^T\boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Z}\right) &= (\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))^T(\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))-\\
&\quad (\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))^T\boldsymbol{D}(\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right)) && \text{(A.76)}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right) &= \mathrm{vec}\left(\boldsymbol{Y}\boldsymbol{L}_c^T-\boldsymbol{F}\boldsymbol{B}\boldsymbol{L}_c^T\right) && \text{(A.77)}\\
\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right) &= \mathrm{vec}\left(\boldsymbol{W}_r\boldsymbol{Y}\boldsymbol{L}_c^T\boldsymbol{W}_c^T-\boldsymbol{W}_r\boldsymbol{F}\boldsymbol{B}\boldsymbol{L}_c^T\boldsymbol{W}_c^T\right) && \text{(A.78)}
\end{aligned}$$

The log-likelihood can be evaluated in $O(NN_{\mathrm{PC}}^2+NN_{\mathrm{PC}}R+NN_{\mathrm{PC}}P+NPR+NN_{\mathrm{PC}}P+NP+NP^2)$ where we only report all quantities depending on N, which are bottleneck for huge sample sizes, and denote in blue all the quantities that have to be computed only once during optimization.

Calculating the gradient The gradient of the likelihood can be written as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_i} = \frac{1}{2}\underbrace{\text{vec}\left(\boldsymbol{Z}\right)^T \boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\text{vec}\left(\boldsymbol{Z}\right)}_{\text{squared form 1}} + \underbrace{\text{vec}\left(\boldsymbol{Z}\right)^T \boldsymbol{K}^{-1}\boldsymbol{X}\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}_i}}_{\text{squared form 2}} \tag{A.79}$$

$$-\frac{1}{2}\underbrace{\text{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right)}_{\text{trace}} - \frac{1}{2}\underbrace{\text{tr}\left(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\theta}_i}\right)}_{\text{reml}} \tag{A.80}$$

Let us start by rewriting $\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}$:

$$\begin{aligned}
\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1} &= \boldsymbol{L}^T\left(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\right)\boldsymbol{L}\left(\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i}\otimes \boldsymbol{R}\right)\boldsymbol{L}^T\left(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\right)\boldsymbol{L} \\
&= \boldsymbol{L}^T\left(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\right)\left(\underbrace{\boldsymbol{L}_c\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}_c^T}_{\tilde{\boldsymbol{C}}}\otimes \boldsymbol{R}\right)\left(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\right)\boldsymbol{L} \\
&= \boldsymbol{L}^T\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{L} + \boldsymbol{L}^T\boldsymbol{W}^T\boldsymbol{D}\left(\underbrace{\boldsymbol{W}_c\tilde{\boldsymbol{C}}\boldsymbol{W}_c^T}_{\bar{\boldsymbol{C}}}\otimes\underbrace{\boldsymbol{W}_r\boldsymbol{R}\boldsymbol{W}_r^T}_{\boldsymbol{S}_r}\right)\boldsymbol{D}\boldsymbol{W}\boldsymbol{L} \\
&\quad -\boldsymbol{L}^T\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L} - \left(\boldsymbol{L}^T\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\right)^T \\
&= \boldsymbol{L}^T\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{L} + \boldsymbol{L}^T\boldsymbol{W}^T\boldsymbol{D}\left(\bar{\boldsymbol{C}}\otimes \boldsymbol{S}_r\right)\boldsymbol{D}\boldsymbol{W}\boldsymbol{L} \\
&\quad -\boldsymbol{L}^T\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L} \\
&\quad -\left(\boldsymbol{L}^T\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\right)^T
\end{aligned} \tag{A.81}$$

where we used that $\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} = \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i}\otimes \boldsymbol{R}$ where $\boldsymbol{C}$ and $\boldsymbol{R}$ are $\boldsymbol{C}_r$ and $\boldsymbol{R}_r$ if $\boldsymbol{\theta}_i$ is a region term parameter or $\boldsymbol{C}_n$ and $\boldsymbol{I}_N$ if $\boldsymbol{\theta}_i$ is a noise term parameter.

The gradients of $\boldsymbol{A}$ and $\boldsymbol{\beta}$ can be calculated as

$$\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\theta}_i} = -\boldsymbol{X}^T\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\boldsymbol{X} \tag{A.82}$$

$$\begin{aligned}
&= -(\boldsymbol{L}\boldsymbol{X})^T\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X} - (\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X})^T\left(\bar{\boldsymbol{C}}\otimes \boldsymbol{S}_r\right)\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X} \\
&\quad +(\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X} + \left((\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X}\right)
\end{aligned} \tag{A.83}$$

$$\begin{aligned}
&= -(\boldsymbol{L}_c^T\tilde{\boldsymbol{C}}\boldsymbol{L}_c)\otimes(\boldsymbol{F}^T\boldsymbol{R}\boldsymbol{F}) - (\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X})^T\left(\bar{\boldsymbol{C}}\otimes \boldsymbol{S}_r\right)\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X} \\
&\quad +(\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X} + \left((\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes \boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X}\right)^T
\end{aligned} \tag{A.84}$$

and

$$\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}_i} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\theta}_i}\boldsymbol{\beta} - \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right) \tag{A.85}$$

where

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right) &= (\boldsymbol{L}\boldsymbol{X})^T\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) \\ &+(\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X})^T\left(\bar{\boldsymbol{C}}\otimes\boldsymbol{S}_r\right)\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) \\ &-(\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X})^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) \\ &-(\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X})^T(\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right)) \end{aligned}\tag{A.86}$$

Several of the matrix products in (A.82, A.85, A.86) have already been computed for estimating the log-likelihood. The additional terms can be computed efficiently by using convenient factorisations and Kronecker product algebra:

$$\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X} = \boldsymbol{W}_{\boldsymbol{c}}\tilde{\boldsymbol{C}}\boldsymbol{L}_c\otimes\boldsymbol{W}_r\boldsymbol{R}\boldsymbol{F} \tag{A.87}$$

$$(\boldsymbol{L}\boldsymbol{X})^T\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) = \mathrm{vec}\left(\boldsymbol{F}\boldsymbol{R}\boldsymbol{Y}\boldsymbol{L}\boldsymbol{c}^T\tilde{\boldsymbol{C}}^T\boldsymbol{L}\boldsymbol{c}\right) \tag{A.88}$$

$$\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) = \mathrm{vec}\left(\boldsymbol{W}_r\boldsymbol{R}\boldsymbol{Y}\boldsymbol{L}_c^T\tilde{\boldsymbol{C}}^T\boldsymbol{W}_{\boldsymbol{c}}^T\right) \tag{A.89}$$

Notice that the computation of $\boldsymbol{RY}$ or $\boldsymbol{RV}$ can also be done in linear time in N. In the non-trivial case where $\boldsymbol{R} = \boldsymbol{G}\boldsymbol{G}^T$ we can rewrite $\boldsymbol{RY} = \boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{Y})$ which takes $O(NRP)$.

The two quadratic terms can be computed respectively as

$$\begin{aligned}\mathrm{vec}\left(\boldsymbol{Z}\right)^T\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Z}\right) &= +(\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))^T\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right)+ \\ &+(\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))^T\left(\bar{\boldsymbol{C}}\otimes\boldsymbol{S}_r\right)\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right)+ \\ &-2(\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right)\end{aligned}\tag{A.90}$$

and

$$\mathrm{vec}\left(\boldsymbol{Z}\right)^T\boldsymbol{K}^{-1}\boldsymbol{X}\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}_i} = (\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))^T\boldsymbol{L}\boldsymbol{X}\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}_i} \tag{A.91}$$

$$-(\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right))^T\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\boldsymbol{X}\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}_i} \tag{A.92}$$

where again many of the matrix products have already been computed while the new terms can also be computed efficiently or rewritten exploiting cached elements. For

example we have

$$\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Z}\right) = \boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) \quad (A.93)$$

$$-\boldsymbol{W}\left(\tilde{\boldsymbol{C}}\otimes\boldsymbol{R}\right)\boldsymbol{L}\boldsymbol{X}\boldsymbol{\beta} \quad (A.94)$$

$$(A.95)$$

Finally, The trace term can be rewritten as

$$\mathrm{tr}\left(\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\right) = \mathrm{tr}\left(\boldsymbol{L}^T\left(\boldsymbol{I}-\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\right)\boldsymbol{L}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\right) \quad (A.96)$$

$$= \mathrm{tr}\left(\boldsymbol{L}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\boldsymbol{L}^T\right)-\mathrm{tr}\left(\boldsymbol{D}\boldsymbol{W}\boldsymbol{L}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\boldsymbol{W}^T\boldsymbol{L}^T\right) \quad (A.97)$$

$$= \mathrm{tr}\tilde{\boldsymbol{C}}\ \mathrm{tr}\boldsymbol{R}-\mathrm{diag}(\boldsymbol{D})^T\mathrm{diag}(\bar{\boldsymbol{C}}\otimes\boldsymbol{S}_r) \quad (A.98)$$

A gradient evaluation takes $O(NRN_{\mathrm{PC}}+NRP+N+NN_{\mathrm{PC}}+NP+NP^2)$ operations, where we just report terms that depend on N as they are bottleneck for large sample size and we report in blue terms that have to computed only once.

A.5 Implementation of mtSet-LowRankBg

Here we consider a low-rank approximation to the background covariance, which leads to a model with two low-rank variance components and a full-rank noise term. Let $\boldsymbol{N}\boldsymbol{N}^T$ be a K rank approximation to the RRM $\boldsymbol{R}_g$. The log-likelihood of the model is

$$\mathcal{L} = \mathrm{const.}-\frac{1}{2}\underbrace{\mathrm{vec}\left(\boldsymbol{Y}\right)^T\boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right)}_{\text{squared form term}}-\frac{1}{2}\underbrace{\log\det\boldsymbol{K}}_{\text{logdet term}} \quad (A.99)$$

where

$$\boldsymbol{K}=\boldsymbol{E}\boldsymbol{E}^T\otimes\boldsymbol{G}\boldsymbol{G}^T+\boldsymbol{C}_g\otimes\boldsymbol{N}\boldsymbol{N}^T+\boldsymbol{C}_n\otimes\boldsymbol{I}_N \quad (A.100)$$

The covariance matrix of the model can be rewritten as

$$\begin{aligned}
\boldsymbol{K} &= \boldsymbol{E}\boldsymbol{E}^T \otimes \boldsymbol{G}\boldsymbol{G}^T + \boldsymbol{C}_g \otimes \boldsymbol{N}\boldsymbol{N}^T + \boldsymbol{C}_n \otimes \boldsymbol{I}_N \\
&= \left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right)\left(\boldsymbol{E}^\star \boldsymbol{E}^{\star T} \otimes \boldsymbol{G}\boldsymbol{G}^T + \boldsymbol{C}_g^\star \otimes \boldsymbol{N}\boldsymbol{N}^T + \boldsymbol{I}_{NP}\right)\left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right)^T \\
&= \left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right)\left(\begin{bmatrix}\boldsymbol{E}^\star \otimes \boldsymbol{G} & \boldsymbol{C}_g^{\star 1/2} \otimes \boldsymbol{N}\end{bmatrix} \underbrace{\begin{bmatrix}\boldsymbol{E}^\star \otimes \boldsymbol{G} & \boldsymbol{C}_g^{\star 1/2} \otimes \boldsymbol{N}\end{bmatrix}^T}_{\boldsymbol{W} \in \mathbb{R}^{(CR+PK)\times NP}} + \boldsymbol{I}_{NP}\right) \times \\
&\quad \left(\boldsymbol{U}_n \boldsymbol{S}_n^{1/2} \otimes \boldsymbol{I}_N\right)^T
\end{aligned} \tag{A.101}$$

and its inverse as

$$\begin{aligned}
\boldsymbol{K}^{-1} &= \left(\boldsymbol{S}_n^{-1/2}\boldsymbol{U}_n^T \otimes \boldsymbol{I}_N\right)^T \left(\boldsymbol{W}^T\boldsymbol{W} + \boldsymbol{I}_{NP}\right)^{-1} \left(\boldsymbol{S}_n^{-1/2}\boldsymbol{U}_n^T \otimes \boldsymbol{I}_N\right) \\
&= \left(\boldsymbol{S}_n^{-1/2}\boldsymbol{U}_n^T \otimes \boldsymbol{I}_N\right)^T \left(\boldsymbol{I}_{NP} - \boldsymbol{W}^T \underbrace{(\boldsymbol{I} + \boldsymbol{W}\boldsymbol{W}^T)}_{\boldsymbol{\Lambda}}{}^{-1}\boldsymbol{W}\right) \times \\
&\quad \underbrace{\left(\boldsymbol{S}_n^{-1/2}\boldsymbol{U}_n^T \otimes \boldsymbol{I}_N\right)}_{\boldsymbol{L}} \qquad \text{(A.102)} \\
&= \boldsymbol{L}^T(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{W})\boldsymbol{L} \qquad \text{(A.103)}
\end{aligned}$$

where

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{I}_{CR} + \boldsymbol{E}^{\star T}\boldsymbol{E}^\star \otimes \boldsymbol{G}^T\boldsymbol{G} & \boldsymbol{E}^{\star T}\boldsymbol{C}_g^{\star 1/2} \otimes \boldsymbol{G}^T\boldsymbol{N} \\ (\boldsymbol{E}^{\star T}\boldsymbol{C}_g^{\star 1/2})^T \otimes (\boldsymbol{G}^T\boldsymbol{N})^T & \boldsymbol{I}_{PK} + \boldsymbol{C}_g^\star \otimes \boldsymbol{N}^T\boldsymbol{N} \end{bmatrix} \in \mathbb{R}^{(CR+PK)\times(CR+PK)} \tag{A.104}$$

and $\boldsymbol{C}_g^{\star 1/2} = \boldsymbol{U}_{\boldsymbol{C}_g^\star}\boldsymbol{S}_{\boldsymbol{C}_g^\star}{}^{-1/2}$.

Evaluating the log-likelihood The squared form can be computed as

$$\begin{aligned}
\mathrm{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{K}^{-1}\mathrm{vec}\left(\boldsymbol{Y}\right) &= \mathrm{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{L}^T\left(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{W}\right)\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) \qquad \text{(A.105)} \\
&= \left(\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right)\right)^T\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) - \left(\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right)\right)^T\boldsymbol{\Lambda}^{-1}\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) \qquad \text{(A.106)}
\end{aligned}$$

where

$$\boldsymbol{W}\boldsymbol{L}\mathrm{vec}\left(\boldsymbol{Y}\right) = \mathrm{vec}\left(\boldsymbol{W}_r\boldsymbol{Y}\boldsymbol{L}_c^T\boldsymbol{W}_c\right). \tag{A.107}$$

While logdet term can be efficiently calculated as

$$\log\det \boldsymbol{K} \quad = \quad N \log\det \boldsymbol{S}_n + \log\det \boldsymbol{\Lambda} \tag{A.108}$$

Reporting only terms that depend on N, a likelihood evaluation takes $O(NR^2 + NRK + NK^2 + NKP + NRP + NP + NP^2)$ where we reported the operations that need to be computed just once in blue.

Calculating the gradient The gradient of the likelihood can be written as

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_i} = \frac{1}{2} \underbrace{\text{vec}\left(\boldsymbol{N}\right)^T \boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}^{-1} \text{vec}\left(\boldsymbol{N}\right)}_{\text{squared form}} - \frac{1}{2} \underbrace{\text{tr}\left(\boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right)}_{\text{trace}} \tag{A.109}$$

$\boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}^{-1}$ can be rewritten as:

$$\begin{aligned}
\boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} \boldsymbol{K}^{-1} \quad &= \quad \boldsymbol{L}^T \left(\boldsymbol{I} - \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W}\right) \boldsymbol{L} \left(\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i} \otimes \boldsymbol{R}\right) \boldsymbol{L}^T \left(\boldsymbol{I} - \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W}\right) \boldsymbol{L} \\
&= \quad \boldsymbol{L}^T \left(\boldsymbol{I} - \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W}\right) \left(\underbrace{\boldsymbol{L}_c \frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i} \boldsymbol{L}_c^T}_{\tilde{\boldsymbol{C}}} \otimes \boldsymbol{R}\right) \left(\boldsymbol{I} - \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W}\right) \boldsymbol{L} \\
&= \quad \boldsymbol{L}^T \left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{L} + \boldsymbol{L}^T \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} (\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}) \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} \boldsymbol{L} \\
&\quad -\boldsymbol{L}^T \left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} \boldsymbol{L} + \left(\boldsymbol{L}^T \left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} \boldsymbol{L}\right)^T \\
&= \quad \boldsymbol{L}^T \left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{L} + \boldsymbol{L}^T \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} (\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}) \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} \boldsymbol{L} \\
&\quad -\boldsymbol{L}^T \left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} \boldsymbol{L} \\
&\quad - \left(\boldsymbol{L}^T \left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{W}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{W} \boldsymbol{L}\right)^T
\end{aligned} \tag{A.110}$$

where

$$\boldsymbol{W}(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}) \boldsymbol{W}^T \quad = \quad \begin{bmatrix} \boldsymbol{E}^{\star T} \otimes \boldsymbol{G}^T \\ \boldsymbol{C}_g^{\star 1/2^T} \otimes \boldsymbol{N}^T \end{bmatrix} (\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}) \begin{bmatrix} \boldsymbol{E}^\star \otimes \boldsymbol{G} & \boldsymbol{C}_g^{\star 1/2} \otimes \boldsymbol{N} \end{bmatrix} = \tag{A.111}$$

$$= \quad \begin{bmatrix} \boldsymbol{E}^{\star T} \tilde{\boldsymbol{C}} \boldsymbol{E}^\star \otimes \boldsymbol{G}^T \boldsymbol{R} \boldsymbol{G} & \boldsymbol{E}^{\star T} \tilde{\boldsymbol{C}} \boldsymbol{C}_g^{\star 1/2} \otimes \boldsymbol{G}^T \boldsymbol{R} \boldsymbol{N} \\ \boldsymbol{C}_g^{\star 1/2^T} \tilde{\boldsymbol{C}} \boldsymbol{E}^\star \otimes \boldsymbol{N}^T \boldsymbol{R} \boldsymbol{G} & \boldsymbol{C}_g^{\star 1/2^T} \tilde{\boldsymbol{C}} \boldsymbol{C}_g^{\star 1/2} \otimes \boldsymbol{N}^T \boldsymbol{R} \boldsymbol{N} \end{bmatrix} \tag{A.112}$$

and we used that $\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}$ can be written as $\frac{\partial \boldsymbol{C}}{\partial \boldsymbol{\theta}_i} \otimes \boldsymbol{R}$ where $\boldsymbol{C}$ and $\boldsymbol{R}$ are $\boldsymbol{C}_r$ and $\boldsymbol{R}_r$ if $\boldsymbol{\theta}_i$ is a region term parameter or $\boldsymbol{C}_n$ and $\boldsymbol{I}_N$ if $\boldsymbol{\theta}_i$ is a noise term parameter.

The quadratic form term can then be computed as

$$\text{vec}\left(\boldsymbol{Y}\right)^T \boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\text{vec}\left(\boldsymbol{Y}\right) = \tag{A.113}$$

$$\left(\boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right)\right)^T \left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right) \tag{A.114}$$

$$+(\boldsymbol{\Lambda}^{-1}\boldsymbol{W}\boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right))^T\boldsymbol{W}(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R})\boldsymbol{W}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{W}\boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right) + \tag{A.115}$$

$$-2\left(\boldsymbol{W}\left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right)\right)^T \boldsymbol{\Lambda}^{-1}\boldsymbol{W}\boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right)$$

where

$$\left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right) = \text{vec}\left(\boldsymbol{R}\boldsymbol{Y}\boldsymbol{L}_c^T\tilde{\boldsymbol{C}}^T\right) \tag{A.116}$$

$$\boldsymbol{W}\left(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R}\right) \boldsymbol{L}\text{vec}\left(\boldsymbol{Y}\right) = \text{vec}\left(\boldsymbol{W}_r\boldsymbol{R}\boldsymbol{Y}\boldsymbol{L}_c^T\tilde{\boldsymbol{C}}^T\boldsymbol{W}_c^T\right) \tag{A.117}$$

while the trace term as

$$\text{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) = \text{tr}\left(\boldsymbol{L}^T\left(\boldsymbol{I} - \boldsymbol{W}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{W}\right)\boldsymbol{L}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) \tag{A.118}$$

$$= \text{tr}\left(\boldsymbol{L}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}^T\right) - \text{tr}\left(\boldsymbol{\Lambda}^{-1}\boldsymbol{W}\boldsymbol{L}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{L}^T\boldsymbol{W}^T\right) \tag{A.119}$$

$$= \text{tr}\tilde{\boldsymbol{C}}\ \text{tr}\boldsymbol{R} - \sum_{ij}\boldsymbol{\Lambda}_{ij}^{-1}(\boldsymbol{W}(\tilde{\boldsymbol{C}} \otimes \boldsymbol{R})\boldsymbol{W}^T)_{ij}. \tag{A.120}$$

Reporting only terms that depend on N, a gradient calculation takes $O(NR^2 + NK^2 + NRK + NKP + NRP + N + NP + NP^2)$ where operations that need to be computed just ones are marked in blue.

A.6 Implementation of iSet for stratification analysis

As we have seen in Section 4.4.1, the iSet model for the analysis of unrelated individuals is

$$\boldsymbol{y} \sim \mathcal{N}\left(\underbrace{\boldsymbol{X}\boldsymbol{\beta}}_{\text{fixed effect covariates}}, \underbrace{\boldsymbol{W}\boldsymbol{W}^T}_{\text{low-rank set component}} + \underbrace{\boldsymbol{D}}_{\text{diagonal noise component}}\right) \tag{A.121}$$

where $\boldsymbol{y} \in \mathbb{R}^N$ is the vector of the observed trait measurements, $\boldsymbol{D}$ is a diagonal matrix with context-specific noise variances and

$$\boldsymbol{W} = \left(\left[\boldsymbol{C}_r^{1/2}\right]_{\boldsymbol{e},:} \otimes \boldsymbol{1}_{1\times R}\right) \odot \left(\boldsymbol{1}_{1\times 2} \otimes \boldsymbol{G}\right), \tag{A.122}$$

where $\boldsymbol{e} \in \{1,2\}^N$ is a context indicator and $\boldsymbol{G} \in \mathbb{R}^{N \times R}$ is the standardised genotype matrix for R variants in the set (see also Section 4.4.1).

LML and gradients

$$\begin{aligned}
\mathcal{L} &= -\frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\log|\boldsymbol{A}| - \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{K}^{-1}\boldsymbol{y} + \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{K}^{-1}\boldsymbol{X}\boldsymbol{\beta} && \text{(A.123)}\\
\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_i} &= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{A}^{-1}\frac{\partial\boldsymbol{A}}{\partial\boldsymbol{\theta}_i}\right) + \frac{1}{2}\boldsymbol{y}^\top\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\boldsymbol{y}\\
&\quad -\boldsymbol{y}^\top\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\boldsymbol{X}\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}^\top\frac{\partial\boldsymbol{A}}{\partial\boldsymbol{\theta}_i}\boldsymbol{\beta} && \text{(A.124)}
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{K} &= \boldsymbol{W}\boldsymbol{W}^T + \boldsymbol{D} && \text{(A.125)}\\
\boldsymbol{A} &= \boldsymbol{X}^\top\boldsymbol{K}^{-1}\boldsymbol{X} && \text{(A.126)}\\
\boldsymbol{\beta} &= \boldsymbol{A}^{-1}\boldsymbol{X}^\top\boldsymbol{K}^{-1}\boldsymbol{y}, && \text{(A.127)}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\boldsymbol{A}}{\partial\boldsymbol{\theta}_i} &= -\boldsymbol{X}^\top\boldsymbol{K}^{-1}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\boldsymbol{X}, && \text{(A.128)}\\
\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i} &= \begin{cases} \boldsymbol{W}\frac{\partial\boldsymbol{W}}{\partial\boldsymbol{\theta}_i}^T + \frac{\partial\boldsymbol{W}}{\partial\boldsymbol{\theta}_i}\boldsymbol{W}^T & \text{if } i \leq \frac{C(C+1)}{2} \\ \frac{\partial\boldsymbol{D}}{\partial\boldsymbol{\theta}_i} \quad \text{otherwise} \end{cases}, && \text{(A.129)}
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{K}^{-1} &= \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{H}^{-1}\boldsymbol{W}^T\boldsymbol{D}^{-1}, && \text{(A.130)}\\
\boldsymbol{H} &= \boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{D}^{-1}\boldsymbol{W} && \text{(A.131)}
\end{aligned}$$

As shown in more detail in the next paragraph, evaluation of a LML and its gradients has complexity $O(NR^2 + NK^2 + R^3 + K^3)$.

Computational complexity of all terms We denote with $\boxed{\cdots}$ the blocks which have been pre-computed at a given stage.

- $\boldsymbol{D}^{-1}\boldsymbol{W}$

$O(NR)$

- $\boldsymbol{H} = \boldsymbol{I} + \boldsymbol{W}^T \boxed{\boldsymbol{D}^{-1}\boldsymbol{W}}$

$O(NR^2)$

- cholesky $(\boldsymbol{H})$

$O(R^3)$

- $\boldsymbol{K}^{-1}\boldsymbol{y} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{H}^{-1}\boldsymbol{W}^T\boldsymbol{D}^{-1}\boldsymbol{y}$

$O(N + NR)$

- $\boldsymbol{K}^{-1}\boldsymbol{X} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}\boldsymbol{W}\boldsymbol{H}^{-1}\boldsymbol{W}^T\boldsymbol{D}^{-1}\boldsymbol{X}$

$O(NK + NRK)$

- $\boldsymbol{A} = \boldsymbol{X}^T \boxed{\boldsymbol{K}^{-1}\boldsymbol{X}}$

$O(NK^2)$

- cholesky $(\boldsymbol{A})$

$O(K^3)$

- $\boldsymbol{\beta} = \boldsymbol{A}^{-1}\boldsymbol{X}^T \boxed{\boldsymbol{K}^{-1}\boldsymbol{y}}$

$O(NK + K^2)$

- $\boxed{\boldsymbol{K}^{-1}\boldsymbol{X}}\boldsymbol{\beta}$

$O(NK)$

- $\log\det \boldsymbol{K} = \log\det(\boldsymbol{H}) - \log\det\left(\boldsymbol{D}^{-1}\right)$

$O(N + R)$

- $\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} \boxed{\boldsymbol{K}^{-1}\boldsymbol{y}} = \begin{cases} \boldsymbol{W}\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}_i}^T \boxed{\boldsymbol{K}^{-1}\boldsymbol{y}} + \frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}_i}\boldsymbol{W}^T \boxed{\boldsymbol{K}^{-1}\boldsymbol{y}} & \text{if } i \leq \frac{E(E+1)}{2} \\ \frac{\partial \boldsymbol{D}}{\partial \boldsymbol{\theta}_i} \boxed{\boldsymbol{K}^{-1}\boldsymbol{y}} & \text{otherwise} \end{cases}$

$O(NR + N)$

- $\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boxed{\boldsymbol{K}^{-1}\boldsymbol{X}} = \begin{cases} \boldsymbol{W}\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}_i}^T\boxed{\boldsymbol{K}^{-1}\boldsymbol{X}} + \frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}_i}\boldsymbol{W}^T\boxed{\boldsymbol{K}^{-1}\boldsymbol{X}} & \text{if } i \leq \frac{E(E+1)}{2} \\ \frac{\partial \boldsymbol{D}}{\partial \boldsymbol{\theta}_i}\boxed{\boldsymbol{K}^{-1}\boldsymbol{X}} & \text{otherwise} \end{cases}$

$O(NRK + NK)$

- $\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{\theta}_i} = -\boxed{\boldsymbol{X}^\top \boldsymbol{K}^{-1}}\boxed{\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\boldsymbol{X}}$

$O(NK^2)$

- $\operatorname{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) = \sum_{ii}\boldsymbol{D}_{ii}^{-1}\left(\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right)_{ii} - \sum_{ij}\boldsymbol{H}_{ij}^{-1}\left(\boldsymbol{W}^T\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{D}^{-1}\boldsymbol{W}\right)_{ij}$
  $= \begin{cases} 2\sum_{ii}\boldsymbol{D}_{ii}^{-1}\left(\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}_i}\boldsymbol{W}^T\right)_{ii} - \sum_{ij}\boldsymbol{H}_{ij}^{-1}\left(\boxed{\boldsymbol{D}^{-1}\boldsymbol{W}}^T\frac{\partial \boldsymbol{W}}{\partial \boldsymbol{\theta}_i}\boxed{\boldsymbol{W}^T\boldsymbol{D}^{-1}\boldsymbol{W}} + (\text{transp})\right)_{ij} \\ \sum_{ii}\boldsymbol{D}_{ii}^{-1}\left(\frac{\partial \boldsymbol{D}}{\partial \boldsymbol{\theta}_i}\right)_{ii} - \sum_{ij}\boldsymbol{H}_{ij}^{-1}\left(\boxed{\boldsymbol{D}^{-1}\boldsymbol{W}}^T\frac{\partial \boldsymbol{D}}{\partial \boldsymbol{\theta}_i}\boxed{\boldsymbol{D}^{-1}\boldsymbol{W}}\right)_{ij} \end{cases}$

$O(NR + N + NR^2 + R^3 + R^2)$

B | Supplementary Results for: Efficient set tests for joint analysis of correlated traits

B.1 Supplementary tables

Dataset	Method	π	a	d
NFBC	mtSet	0.06	0.55	2.67
	mtSet-PC	0.06	0.50	3.12
	stSet (crp3dec)	0.65	0.50	0.94
	stSet (FS_KOL_L)	0.64	0.50	0.89
	stSet (FS_KOL_H)	0.65	0.50	0.89
	stSet (FS_TRIGL)	0.65	0.50	0.94
Rat dataset	mtSet	0.22	0.59	2.08
	mtSet-PC	0.19	0.55	2.33
	stSet (basos)	0.91	0.50	0.59
	stSet (eos)	0.93	0.35	1.14
	stSet (lucs)	0.92	0.50	0.74
	stSet (lymphs)	0.92	0.35	1.04
	stSet (monos)	0.93	0.59	0.55
	stSet (neuts)	0.93	0.35	1.14

Table B.1: **Estimated parameters of the parametric LLR distribution.** The P values for the proposed set test are obtained from a log-likelihood ratio (LLR) test statistics, assuming $\frac{1}{a}\text{LLR} \sim \pi\chi^2_0 + (1-\pi)\chi^2_d$. The mixture ($\pi$), the scale ($a$) and the degree-of-freedom (d) parameters are estimated by fitting the parametric form to the distribution of LLRs obtained through permutations (see Section 3.1.2 for more details). We here report the estimated values of π, a and d for mtSet, mtSet-PC and the single-trait analyses with stSet for the two real-data applications presented in Chapter 3.

Inverse		
$\boldsymbol{A}^{-1}$		$O(N^3 + P^3)$
Cholesky		
$\text{chol}(\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W})$	$\boldsymbol{D}_\wedge\boldsymbol{W}$ ${\boldsymbol{W}^T}_\wedge(\boldsymbol{D}\boldsymbol{W})$ $\text{chol}(\boldsymbol{I} + \boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W})$	$O(NPCR)$ $O(NPRC^2 + NR^2C^2)$ $O(C^3R^3)$
Log Likelihood		
$\log\det\boldsymbol{K}$	$\log\det\boldsymbol{A}$ $\log\det\boldsymbol{\Lambda}$	$O(NP)$ $O(CR)$
$\tilde{\boldsymbol{y}}^T\boldsymbol{D}\tilde{\boldsymbol{y}} - \bar{\boldsymbol{y}}^T\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{y}}$	$\tilde{\boldsymbol{y}} = \boldsymbol{L}\text{vec}(\boldsymbol{Y})$ $\bar{\boldsymbol{y}} = \boldsymbol{W}^T\boldsymbol{D}\text{vec}(\boldsymbol{Y})$ $\tilde{\boldsymbol{y}}^T\boldsymbol{D}\tilde{\boldsymbol{y}} - \bar{\boldsymbol{y}}^T\boldsymbol{\Lambda}^{-1}\bar{\boldsymbol{y}}$	$O(N^2P + NP^2)$ $O(NP + NPC + NRC)$ $O(NP + C^2R^2 + CR)$
Gradient		
$\widetilde{\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}} = \boldsymbol{L}\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}\boldsymbol{L}^T$		$O(N^2R + P^2C)$ *region* $O(N^3 + P^3)$ *rand effect* $O(N^3 + P^3)$ *noise*
$\widetilde{\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}} = \boldsymbol{W}^T\boldsymbol{D}\widetilde{\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}}\boldsymbol{D}\boldsymbol{W}$	$\widetilde{\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}}_\wedge(\boldsymbol{D}\boldsymbol{W})$	$O(NRPC^2 + NR^2C^2)$ *region* $O(NRPC + NRP^2C)$ *rand eff* $O(NRP^2C)$ *noise*
	${\boldsymbol{W}^T}_\wedge(\boldsymbol{D}\widetilde{\frac{\partial\boldsymbol{K}}{\partial\boldsymbol{\theta}_i}}\boldsymbol{D}\boldsymbol{W})$	$O(NPRC^2 + NR^2C^2)$
**computed only once		
**computed only once per-region		

Table B.2: **Tabular summary of the complexity of individual computational steps in the mtSet inference.**

h_r^2	0.001	0.002	0.004	**0.006**	0.008	0.010	-	-
S_r	1	2	5	**8**	12	20	-	-
π_r	0.00	0.125	0.375	**0.50**	0.75	1.00	-	-
α	0.0	0.125	0.375	**0.50**	0.75	1.00	-	-
h_g^2	0.00	0.10	0.20	**0.40**	0.60	0.80	-	-
λ	0.00	0.10	0.20	0.40	**0.60**	0.80	-	-
window size (in kb)	1	5	10	20	**30**	50	100	200

Table B.4: **Parameter ranges for simulated datasets.** To assess the power of different methods, we considered a range of alternative simulations, varying key parameters that determine the genetic architecture of the traits. We altered the variance explained by the region (h_r^2), the number of causal variants from the region (S_r), the percentage of shared causal variants (π_r), the percentage of background and residual signal that is shared across traits (α), the variance explained by genetic background (h_g^2), the percentage of residual variance explained by hidden confounders (λ) and the window size. Each of those parameters was varied while keeping the other values at the default value (highlighted in bold).

		Single Trait ($P = 1$) - Gradient-free		Multi Trait ($P > 1$) - Gradient-based	
Test type	**Structure of $\mathbf{C_g}$**	**Method**	**Per-test complexity**	**Method**	**Per-test complexity ($P > 1$)**
Single variant test	full-rank	Fast-LMM, Gemma	$O(N^2 + tN)$	mvLMM	$O(N^2 + t_1 NP^2 + t_2 NP^6)$
Set test	-	mtSet-PC	$O(N(R + N_{\text{PC}})^2 + tNN_{\text{PC}}$	mtSet-PC	$O(NR^2 + N_{\text{PC}}^2 + NRP + tNP^2(N_{\text{PC}} + P^2))$
Set test	low-rank	FaST-LMM-Set	$O(N(R + K)^2 + tN)$	mtSet-LowRankBg	$O(NR^2 + NK^2 + NRP + NKP + tNP^4)$
Set test	full-rank	stSet	$O(N^2R + tNR)$	mtSet	$O(N^2R + t(NR^2P^2 + NRP^4))$

Table B.3: **Comparison of the mtSet inference method and other linear mixed models implementations used in GWAS for testing associations.** For all methods with exception of mtSet-PC we assume that the number of covariates is 1. N denotes the number of samples, P the number of traits and t is the number of iterations of the algorithm (unidimensional Brenth search for efficient LMM and FaST-LMM-Set, LBFGS for mtSet, mtSet-PC and mtSet-LowRankBg, expectation maximization and Newton-Raphson for mvLMM respectively denoted as t_1 and t_2), R is the number of variants in the sets, N_{PC} the number of principal components used in mtSet-PC to model population structure and K is the rank of the low-rank approximation of the genetic relatedness matrix used in FaST-LMM-Set and mtSet-LowRankBg.

Simulated term	**Variance explained**
reg shared	$\pi_r h_r^2$
reg ind	$(1-\pi_r)h_r^2$
pop struct shared	αh_g^2
pop struct ind	$(1-\alpha)h_g^2$
noise conf shared	$\lambda\alpha(1-h_g^2-h_r^2)$
noise conf ind	$\lambda(1-\alpha)(1-h_g^2-h_r^2)$
noise iid	$(1-\lambda)(1-\alpha)(1-h_g^2-h_r^2)$

Table B.5: **Average variance across traits explained by each term in the simulation framework.** The table shows how the variance explained by each of the simulated terms is related to the simulation parameters. The simulation parameters are the variance explained by the causal region (h_r^2), the number of causal variants in the region (S_r), the percentage of shared causal variants (π_r), the variance explained by relatedness effects (h_g^2), the fraction of residual variance that is not iid across samples (λ), and the fraction of relatedness and residual signal that is shared across traits (α).

Method	Dataset	Significance Level	True Windows	Test Windows	Train Windows
mtSet	1000G	α =5.00e-02	4.25e-02	5.41e-02	4.89e-02
	1000G	α =5.00e-03	4.26e-03	5.62e-03	5.08e-03
	1000G	α =5.00e-04	4.33e-04	5.73e-04	4.94e-04
	1000G	α =5.00e-05	5.37e-05	4.81e-05	5.39e-05
mtSet-PC	1000G	α =5.00e-02	5.41e-02	4.83e-02	4.95e-02
	1000G	α =5.00e-03	6.49e-03	4.82e-03	5.08e-03
	1000G	α =5.00e-04	7.41e-04	4.13e-04	5.05e-04
	1000G	α =5.00e-05	1.09e-04	3.47e-05	4.68e-05
mtSet	simPopStructure	α =5.00e-02	4.70e-02	4.75e-02	4.92e-02
	simPopStructure	α =5.00e-03	4.87e-03	4.45e-03	5.00e-03
	simPopStructure	α =5.00e-04	4.53e-04	3.92e-04	4.77e-04
	simPopStructure	α =5.00e-05	4.92e-05	2.91e-05	4.01e-05
mtSet-PC	simPopStructure	α =5.00e-02	5.09e-02	4.86e-02	5.06e-02
	simPopStructure	α =5.00e-03	5.14e-03	4.40e-03	4.91e-03
	simPopStructure	α = 5.00e-04	5.17e-04	3.95e-04	4.66e-04
	simPopStructure	α =5.00e-05	4.03e-05	2.46e-05	4.23e-05
mtSet	simUnrelated	α =5.00e-02	4.57e-02	4.55e-02	4.79e-02
	simUnrelated	α =5.00e-03	4.56e-03	4.63e-03	4.96e-03
	simUnrelated	α =5.00e-04	4.06e-04	4.43e-04	4.72e-04
	simUnrelated	α =5.00e-05	5.37e-05	3.13e-05	5.06e-05
mtSet-PC	simUnrelated	α =5.00e-02	5.27e-02	4.89e-02	5.01e-02
	simUnrelated	α =5.00e-03	5.32e-03	4.89e-03	4.99e-03
	simUnrelated	α =5.00e-04	5.03e-04	4.62e-04	4.62e-04
	simUnrelated	α =5.00e-05	6.72e-05	2.69e-05	4.54e-05
mtSet	simRelated	α =5.00e-02	4.39e-02	4.72e-02	5.02e-02
	simRelated	α =5.00e-03	4.14e-03	4.32e-03	4.94e-03
	simRelated	α =5.00e-04	4.22e-04	4.06e-04	4.48e-04
	simRelated	α =5.00e-05	2.24e-05	4.48e-05	3.11e-05
mtSet-PC	simRelated	α =5.00e-02	1.40e-01	4.82e-02	5.00e-02
	simRelated	α =5.00e-03	2.86e-02	4.98e-03	5.19e-03
	simRelated	α =5.00e-04	5.57e-03	4.78e-04	5.22e-04
	simRelated	α =5.00e-05	9.77e-04	4.70e-05	4.45e-05

Table B.6: **Type-1 error estimates on simulated data.** Shown are the type-1 error estimates for increasingly stringent α level thresholds $\alpha \in \{0.05, 0.005, 0.005, 0.0005\}$ on four alternative simulated datasets (see also **Fig. B.3**). Train windows denote regions that have been used (based on permutations) to fit the parametric model of null distribution. True windows denote genomic regions that have not been used to train the null model (independent test validation). Finally, test windows denote regions where the genotype and phenotype relationship have been shuffled. These are equivalent to train windows, but using a different set of permutations. mtSet and mtSet-PC perform equally well when no structure or population structure is present, while the calibration of mtSet-PC detoriates when the individuals are related.

	CRP	LDL	HDL	TRIGL
Heritability Estimates				
single-trait	0.11±0.04	0.37±0.02	0.36±0.02	0.15±0.04
multi-trait	0.11±0.04	0.36±0.02	0.36±0.02	0.16±0.03
Genetic Covariance Matrix				
CRP	0.11±0.05	-0.03±0.03	0.06±0.03	-0.10±0.04
LDL	-0.03±0.03	0.36±0.05	-0.05±0.04	0.06±0.04
HDL	0.06±0.03	-0.05±0.04	0.36±0.05	-0.11±0.04
TRIGL	-0.10±0.04	0.06±0.04	-0.11±0.04	0.16±0.05
Noise Covariance Matrix				
CRP	0.89±0.05	0.13±0.03	-0.24±0.04	0.36±0.04
LDL	0.13±0.03	0.63±0.05	-0.07±0.04	0.28±0.04
HDL	-0.24±0.04	-0.07±0.04	0.64±0.05	-0.25±0.04
TRIGL	0.36±0.04	0.28±0.04	-0.25±0.04	0.84±0.05
Phenotypic Covariance Matrix				
CRP	1.00±0.00	0.10±0.01	-0.18±0.01	0.26±0.01
LDL	0.10±0.01	1.00±0.00	-0.12±0.01	0.34±0.01
HDL	-0.18±0.01	-0.12±0.01	1.00±0.00	-0.35±0.01
TRIGL	0.26±0.01	0.34±0.01	-0.35±0.01	1.00±0.00

Table B.7: **Estimates of trait heritability and covariances for 4 lipid-related traits from the NFBC dataset.**
Heritability estimates: Single-trait heritability estimates are obtained independently for each trait. Multi-trait estimates correspond to the marginal estimates obtained form the genetic and noise trait covariance matrix from the null model fit of mtSet. As expected, these marginal estimates are consistent.
Genetic covariance matrix: Trait-trait covariances of the relatedness component from the null model fit of mtSet.
Noise covariance matrix: Trait-trait covariances of the noise component of the null model fit of mtSet.
Phenotype covariance: Empirical covariance matrix of the raw phenotypes. All estimates are obtained from a maximum likelihood fit of mtSet; standard errors are denoted by ±.

chr	pos	Set tests		stSet					Single-variant tests		stLMM-SV		
		mtSet	mtSet-PC	CRP	LDL	HDL	TRIG	pos1	mtLMM-SV	CRP	LDL	HDL	TRIG
PCSK9[1]													
1	55250000	**2.0E-12**	**1.8E-12**	3.6E-01	**3.8E-14**	3.6E-01	3.5E-01	-	-	-	-	-	-
1	55300000	**1.3E-14**	**2.2E-14**	3.6E-01	**1.6E-15**	3.6E-01	3.5E-01	-	-	-	-	-	-
1	55450000	2.8E-06	5.2E-06	3.6E-01	**3.8E-08**	3.6E-01	2.1E-01	-	-	-	-	-	-
1	55500000	**2.8E-08**	**1.1E-08**	2.9E-01	**2.8E-10**	3.6E-01	2.7E-01	-	-	-	-	-	-
ANGPTL3[1]													
1	62850000	**1.4E-07**	**9.4E-08**	3.6E-01	3.0E-05	2.7E-02	7.4E-04	-	-	-	-	-	-
SORT1[1], CELSR2[2]													
1	109600000	**1.6E-13**	**3.1E-14**	4.0E-02	1.1E-13	5.0E-02	1.1E-01	109620053	**6.4E-16**	1.2E-01	**2.9E-15**	7.3E-02	1.0E+00
1	109650000	**2.2E-13**	6.2E-06	1.6E-01	1.2E-13	9.0E-02	2.3E-01	-	-	-	-	-	-
CRP[2]													
1	157850000	**2.4E-11**	**4.8E-12**	**2.0E-13**	2.4E-02	3.6E-01	8.5E-02	157908973	**4.8E-14**	**5.3E-16**	9.3E-01	8.6E-02	5.0E-01
1	157900000	**2.7E-23**	**1.7E-25**	**3.0E-25**	1.5E-01	2.0E-01	3.5E-01	157914612	**7.0E-13**	**3.0E-14**	5.7E-01	8.1E-01	9.6E-01
1	157950000	**8.4E-25**	**3.0E-27**	**5.0E-27**	3.6E-01	1.8E-01	3.5E-01	157919563	**6.5E-16**	**4.9E-18**	9.5E-01	4.6E-02	4.1E-01
1	158000000	**1.3E-14**	**6.6E-16**	**3.0E-16**	3.6E-01	3.6E-01	3.5E-01	157945440	**3.6E-21**	**1.3E-22**	7.1E-01	2.8E-01	8.9E-01
1	-	-	-	-	-	-	-	157966663	**1.5E-08**	**5.8E-09**	1.5E-01	6.0E-02	7.1E-01
APOB[1,2]													
2	21000000	2.3E-07	**6.3E-09**	2.7E-01	4.6E-09	4.6E-04	1.8E-03	21047434	**1.1E-08**	1.2E-01	6.8E-07	4.0E-06	8.8E-06
2	21050000	**5.0E-08**	**9.0E-10**	2.4E-01	1.4E-08	5.5E-05	1.1E-04	21059688	**1.1E-08**	1.0E-01	5.4E-07	4.7E-06	1.2E-05
2	21100000	**4.2E-09**	**6.3E-11**	3.6E-01	1.0E-10	1.4E-04	1.9E-04	21085700	1.3E-07	4.1E-01	**1.8E-09**	3.0E-02	7.6E-03
2	21150000	**8.1E-09**	**3.0E-09**	3.6E-01	1.8E-11	4.6E-02	3.5E-01	21091049	**2.5E-08**	1.6E-01	6.5E-06	2.4E-06	3.8E-06
2	21200000	**1.8E-07**	**1.0E-07**	3.6E-01	5.1E-10	3.8E-02	3.5E-01	21165046	**5.1E-08**	7.3E-01	**5.3E-10**	1.3E-01	1.1E-01
GCKR[1,2]													
2	27550000	1.8E-06	**1.6E-07**	3.6E-01	2.4E-02	3.6E-01	**1.8E-08**	27584444	**2.1E-08**	2.9E-02	1.1E-01	1.7E-01	**3.5E-10**
2	27600000	7.1E-07	**5.7E-08**	7.7E-02	1.0E-02	3.6E-01	**6.3E-09**	27594741	2.3E-07	7.5E-02	2.6E-01	3.5E-01	**6.2E-09**
PPP1R3B[1,2]													
8	9200000	**6.9E-08**	**4.8E-09**	2.1E-03	6.9E-03	1.3E-03	3.0E-01	9215142	**1.6E-09**	6.0E-04	2.4E-03	2.8E-05	9.2E-01
8	9250000	**1.1E-07**	**9.5E-09**	4.5E-03	1.1E-02	1.3E-03	3.5E-01	9222556	**3.8E-10**	1.5E-03	6.9E-04	1.5E-05	5.3E-01
LPL[1,2]													
8	19850000	**6.1E-08**	**5.0E-08**	2.1E-01	3.6E-01	1.3E-04	8.4E-07	19875201	**7.8E-10**	2.9E-01	9.0E-01	2.9E-06	**4.6E-09**
8	19900000	**5.5E-12**	**2.5E-12**	3.6E-01	3.6E-01	3.0E-07	**8.0E-09**	-	-	-	-	-	-
8	19950000	**7.7E-12**	**8.7E-12**	2.5E-01	3.6E-01	**2.2E-08**	**7.1E-09**	-	-	-	-	-	-
FADS[1,2]													
11	61300000	7.0E-07	**1.8E-08**	2.9E-01	8.7E-07	5.4E-03	9.6E-03	61314379	**2.2E-08**	8.1E-01	3.3E-06	2.8E-02	1.4E-02
11	61350000	2.8E-07	**2.6E-09**	3.6E-01	4.8E-06	2.4E-02	6.2E-02	61326406	**1.7E-08**	7.9E-01	2.4E-06	5.0E-02	1.2E-02
11	61400000	**1.1E-07**	**1.7E-09**	3.6E-01	2.3E-06	3.4E-03	5.1E-02	-	-	-	-	-	-
APOA1-C3-A4-A5[1]													
11	116100000	7.7E-07	6.2E-07	3.6E-01	2.3E-02	6.0E-03	**4.6E-08**	-	-	-	-	-	-
11	116150000	**2.4E-08**	**2.8E-08**	3.6E-01	5.7E-02	2.7E-03	**8.7E-10**	-	-	-	-	-	-
HNF1A[1,2]													
12	119850000	**4.9E-10**	**2.7E-11**	**4.9E-12**	3.6E-01	3.6E-01	1.5E-01	119873345	**5.1E-12**	**1.3E-13**	9.4E-01	1.8E-01	8.9E-01
12	119900000	**6.1E-11**	**2.6E-12**	**1.1E-12**	3.6E-01	1.9E-01	2.0E-01	119888107	**3.0E-11**	**7.2E-13**	9.3E-01	1.4E-01	7.9E-01
12	119950000	**3.6E-08**	4.1E-01	**2.1E-08**	3.6E-01	3.6E-01	3.5E-01	119915608	**6.8E-10**	**2.2E-10**	3.7E-01	4.7E-01	3.1E-01
12	-	-	-	-	-	-	-	119919810	**7.9E-10**	**1.8E-10**	5.1E-01	4.1E-01	3.5E-01
12	-	-	-	-	-	-	-	119923227	**7.1E-09**	**5.1E-09**	3.1E-01	5.7E-01	1.9E-01
LIPC[1,2]													
15	56450000	**1.1E-12**	**1.4E-14**	3.6E-01	9.1E-02	**1.6E-09**	4.4E-02	56468097	**1.4E-09**	4.8E-01	3.5E-02	1.0E-07	2.3E-01
15	56500000	**4.4E-24**	**7.6E-26**	1.2E-01	2.8E-01	**6.7E-15**	7.8E-04	56470658	**8.2E-16**	6.7E-01	3.4E-01	**6.6E-13**	4.0E-02
15	56550000	**1.7E-12**	**1.2E-12**	1.8E-01	3.6E-01	4.1E-07	4.3E-03	56478046	**1.0E-09**	5.4E-01	5.7E-01	1.8E-08	9.0E-02
15	-	-	-	-	-	-	-	56524633	**8.4E-09**	6.1E-02	4.7E-01	1.4E-03	4.6E-05
15	-	-	-	-	-	-	-	56529710	**4.6E-09**	6.3E-02	2.1E-01	5.7E-04	6.4E-05
CETP[1,2]													
16	55500000	**1.0E-08**	**3.1E-09**	2.2E-01	3.6E-01	**1.9E-10**	3.5E-01	55542640	**7.0E-09**	7.0E-01	7.9E-01	**7.2E-10**	9.5E-01
16	55550000	**1.9E-33**	**3.6E-36**	4.8E-02	1.8E-01	**5.5E-35**	3.2E-01	55550825	**8.7E-36**	2.7E-01	7.2E-02	**4.4E-36**	2.1E-01
16	55600000	**1.7E-33**	**2.8E-36**	8.3E-02	2.1E-01	**4.6E-35**	3.5E-01	55562980	**1.8E-24**	1.4E-01	1.8E-01	**3.6E-26**	7.1E-02
16	-	-	-	-	-	-	-	55564091	**1.4E-19**	2.7E-01	8.0E-01	**5.8E-19**	5.9E-01
LCAT[1,2]													
16	66550000	2.1E-06	**3.8E-08**	3.6E-01	3.2E-01	7.8E-08	3.2E-01	-	-	-	-	-	-
LDLR[1,2]													
19	11050000	**2.0E-09**	**4.6E-10**	2.6E-02	5.4E-12	2.7E-01	2.5E-01	-	-	-	-	-	-
19	11100000	**6.1E-08**	**1.7E-08**	1.4E-01	2.9E-10	3.6E-01	1.4E-01	-	-	-	-	-	-
APOE-C1-C2[1,2]													
19	49850000	**1.9E-08**	**1.2E-08**	8.8E-03	6.4E-06	1.6E-01	3.5E-01	50087106	**1.0E-08**	2.2E-01	**5.1E-09**	1.5E-02	1.4E-03
19	49900000	**1.1E-10**	**7.6E-12**	9.2E-04	6.7E-10	1.2E-01	1.2E-02	50087459	**4.7E-12**	9.9E-06	1.1E-05	4.0E-02	1.2E-04
19	49950000	**1.8E-09**	**5.7E-11**	3.7E-02	4.2E-10	3.6E-01	1.2E-02	-	-	-	-	-	-
19	50050000	**3.4E-19**	**1.7E-21**	7.1E-08	1.5E-16	1.4E-01	1.9E-03	-	-	-	-	-	-
19	50100000	**2.1E-32**	**8.3E-36**	**4.1E-12**	1.0E-25	1.5E-01	8.2E-05	-	-	-	-	-	-
19	50150000	**1.4E-20**	**1.6E-22**	**3.7E-12**	2.1E-06	3.6E-01	7.0E-03	-	-	-	-	-	-

Table B.8: **Tabular summary of QTLs identified by different set tests and single-variant LMMs on the NFBC dataset.** The table shows all significant associations ($\alpha < 0.01$, Bonferroni adjusted) found by different set test (left) and single-variant LMMs (right) grouped by locus. Significant associations ($\alpha < 0.01$, Bonferroni adjusted) are boldified. The numbers in square brackets after the gene names indicate whether the locus was identified in Teslovich et al. (2010), [1], in Zhou and Stephens (2014), [2], or both, [1,2].

Region	mtSet (100 kb)			mtSet (60 kb)			mtSet (300 kb)		
	chrom	pos	pv	chrom	pos	pv	chrom	pos	pv
PCSK9[1]	1	55250000	**1.95E-12**	1	55260000	**1.01E-11**	1	55200000	**7.88E-13**
	1	55300000	**1.30E-14**	1	55290000	**5.04E-09**	1	55350000	**1.77E-17**
	1	55500000	**2.84E-08**	1	55320000	**2.64E-08**	1	55500000	**3.44E-09**
				1	55500000	**5.60E-08**			
ANGPTL3[1]	1	62850000	**1.44E-07**	1	62940000	7.91E-04	1	62850000	1.21E-06
SORT1[1], CELSR2[2]	1	109600000	**1.59E-13**	1	109590000	**1.28E-11**	1	109500000	**1.81E-12**
	1	109650000	**2.23E-13**	1	109620000	**2.75E-15**	1	109650000	**7.85E-13**
				1	109650000	**1.29E-14**			
CRP[2]	1	157850000	**2.44E-11**	1	157860000	**4.27E-11**	1	157800000	**2.20E-25**
	1	157900000	**2.71E-23**	1	157890000	**1.76E-21**	1	157950000	**1.46E-23**
	1	157950000	**8.44E-25**	1	157920000	**4.26E-25**	1	158100000	**5.06E-11**
	1	158000000	**1.25E-14**	1	157950000	**1.07E-28**			
				1	157980000	**7.61E-17**			
				1	158010000	**2.44E-09**			
APOB[1,2]	2	21050000	**5.00E-08**	2	21030000	**3.33E-08**	2	21000000	**2.94E-10**
	2	21100000	**4.19E-09**	2	21060000	**1.19E-08**	2	21150000	**1.14E-11**
	2	21150000	**8.09E-09**	2	21090000	**1.27E-07**			
	2	21200000	**1.77E-07**	2	21120000	**1.58E-09**			
				2	21150000	**2.36E-09**			
				2	21180000	**4.12E-08**			
PPP1R3B[1,2]	8	9200000	**6.87E-08**	8	9210000	**4.98E-09**	8	9150000	**7.34E-08**
	8	9250000	**1.09E-07**	8	9240000	**7.47E-09**	8	9300000	**4.81E-07**
LPL[1,2]	8	19850000	**6.12E-08**	8	19860000	**8.58E-09**	8	19800000	**4.54E-09**
	8	19900000	**5.55E-12**	8	19890000	**6.56E-10**	8	19950000	**1.97E-09**
	8	19950000	**7.71E-12**	8	19920000	**7.65E-15**			
FADS[1,2]	11	61400000	**1.12E-07**	11	61350000	**6.09E-08**	11	61350000	**4.98E-07**
APOA1-C3-A4-A5[1]	11	116150000	**2.42E-08**	11	116130000	**1.93E-09**	11	116250000	8.97E-07
				11	116160000	**8.27E-09**			
HNF1A[1,2]	12	119850000	**4.93E-10**	12	119850000	**4.04E-10**	12	119850000	**1.40E-11**
	12	119900000	**6.05E-11**	12	119880000	**4.35E-11**	12	120000000	**5.07E-10**
	12	119950000	**3.61E-08**	12	119910000	**2.92E-11**			
				12	119940000	**1.77E-09**			
LIPC[1,2]	15	56450000	**1.10E-12**	15	56460000	**2.22E-14**	15	56400000	**5.56E-28**
	15	56500000	**4.35E-24**	15	56490000	**1.44E-26**	15	56550000	**9.26E-27**
	15	56550000	**1.71E-12**	15	56520000	**5.40E-13**			
				15	56550000	**2.35E-09**			
CETP[1,2]	16	55500000	**1.02E-08**	16	55530000	**1.51E-33**	16	55350000	**2.02E-07**
	16	55550000	**1.94E-33**	16	55560000	**1.64E-38**	16	55500000	**2.37E-32**
	16	55600000	**1.70E-33**	16	55590000	**3.43E-29**	16	55650000	**1.58E-31**
LDLR[1,2]	19	11050000	**1.97E-09**	19	11070000	**6.31E-11**	19	10950000	**8.15E-10**
	19	11100000	**6.06E-08**	19	11100000	**5.47E-08**	19	11100000	**2.02E-08**
	19	49850000	**1.90E-08**	19	49920000	**3.33E-10**	19	49800000	**8.48E-11**
APOE-C1-C2[1,2]	19	49900000	**1.05E-10**	19	49950000	**4.65E-10**	19	49950000	**6.04E-26**
	19	49950000	**1.82E-09**	19	50070000	**1.84E-19**	19	50100000	**3.36E-32**
	19	50050000	**3.39E-19**	19	50100000	**3.49E-36**	19	50250000	**4.88E-16**
	19	50100000	**2.11E-32**	19	50130000	**7.73E-23**			
	19	50150000	**1.37E-20**						

Table B.9: **Tabular summary of QTLs identified by mtSet with varying window size on the NFBC dataset.** The table shows all significant associations ($\alpha < 0.01$, Bonferroni adjusted) found by mtSet when considering different window sizes. Specifically, we considered window sizes of 60kb, 100kb and 300kb. The results are overall robust, with only the *PCSK9* locus missed by both the 60kb and the 300kb analysis, and the *APOA1-C3-A4-A5* locus missed by the 300kb analysis.

	basos	eos	lucs	monos	neuts
Heritability Estimates					
single-trait	0.29±0.03	0.44±0.03	0.33±0.03	0.59±0.02	0.46±0.03
multi-trait	0.31±0.03	0.44±0.03	0.33±0.03	0.60±0.02	0.46±0.03
Genetic Covariance Matrix					
basos	0.31±0.05	0.14±0.04	0.22±0.04	0.34±0.05	0.21±0.04
eos	0.14±0.04	0.49±0.07	0.04±0.04	0.14±0.05	0.20±0.05
lucs	0.22±0.04	0.04±0.04	0.34±0.05	0.39±0.05	0.22±0.04
monos	0.34±0.05	0.14±0.05	0.39±0.05	0.66±0.07	0.26±0.05
neuts	0.21±0.04	0.20±0.05	0.22±0.04	0.26±0.05	0.48±0.06
Noise Covariance Matrix					
basos	0.68±0.03	0.18±0.03	0.21±0.03	0.31±0.03	0.27±0.03
eos	0.18±0.03	0.64±0.04	0.17±0.03	0.19±0.02	0.19±0.03
lucs	0.21±0.03	0.17±0.03	0.70±0.03	0.27±0.02	0.30±0.03
monos	0.31±0.03	0.19±0.02	0.27±0.02	0.45±0.03	0.26±0.02
neuts	0.27±0.03	0.19±0.03	0.30±0.03	0.26±0.02	0.56±0.03
Phenotypic Covariance Matrix					
basos	1.00±0.00	0.28±0.03	0.42±0.02	0.62±0.02	0.48±0.02
eos	0.28±0.03	1.00±0.00	0.20±0.03	0.30±0.03	0.33±0.03
lucs	0.42±0.02	0.20±0.03	1.00±0.00	0.60±0.02	0.51±0.02
monos	0.62±0.02	0.30±0.03	0.60±0.02	1.00±0.00	0.50±0.02
neuts	0.48±0.02	0.33±0.03	0.51±0.02	0.50±0.02	1.00±0.00

Table B.10: **Estimates of trait heritability and covariances for 6 phenotypes related to basal haematology on the rat dataset.**
Heritability estimates: Single-trait heritability estimates are obtained independently for each trait. Multi-trait estimates correspond to the marginal estimates obtained form the genetic and noise trait covariance matrix from the null model fit of mtSet. As expected, these marginal estimates are consistent.
Genetic covariance matrix: Trait-trait covariances of the relatedness component from the null model fit of mtSet.
Noise covariance matrix: Trait-trait covariances of the noise component of the null model fit of mtSet.
Phenotype covariance: Empirical covariance matrix of the raw phenotypes. All estimates are obtained from a maximum likelihood fit of mtSet; standard errors are denoted by ±.

	Candidate Windows	mtSet	stSet					
			basos	eos	lucs	lymphs	monos	neuts
1	*chrom1:273500000*	**9.76E-07**	9.27E-02	7.18E-02	1.12E-02	8.15E-02	1.71E-05	7.46E-02
2	*chrom1:274000000*	**4.65E-07**	9.27E-02	7.18E-02	2.02E-02	8.15E-02	8.45E-06	7.46E-02
3	*chrom8:128500000*	**4.99E-07**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	7.63E-06	7.46E-02
4	*chrom8:129000000*	**1.56E-07**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	8.17E-06	7.46E-02
5	*chrom8:129500000*	**1.11E-07**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	9.34E-06	7.46E-02
6	*chrom8:130000000*	**3.23E-07**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	1.31E-05	7.46E-02
7	*chrom8:130500000*	**1.17E-08**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	3.29E-07	7.46E-02
8	*chrom8:131000000*	**2.17E-09**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	**5.15E-08**	7.46E-02
9	*chrom8:131500000*	**1.27E-09**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	**2.78E-08**	7.46E-02
10	*chrom8:132000000*	**1.21E-09**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	**3.71E-08**	7.46E-02
11	*chrom8:132500000*	**1.06E-09**	9.27E-02	7.18E-02	8.12E-02	8.15E-02	**4.68E-08**	7.46E-02
12	*chrom9:500000*	**7.99E-10**	9.27E-02	7.18E-02	8.12E-02	**8.84E-13**	7.19E-02	7.46E-02
13	*chrom9:1000000*	**2.05E-10**	3.99E-03	7.18E-02	1.88E-03	**1.67E-14**	7.19E-02	7.46E-02
14	*chrom9:1500000*	**6.55E-11**	1.83E-03	7.18E-02	1.17E-03	**1.04E-14**	7.19E-02	7.46E-02
15	*chrom9:2000000*	**6.20E-10**	7.99E-03	7.18E-02	1.13E-02	**1.22E-11**	7.19E-02	7.46E-02
16	*chrom9:2500000*	**3.38E-09**	1.05E-02	7.18E-02	8.12E-02	**3.85E-11**	7.19E-02	7.46E-02
17	*chrom9:3000000*	**8.66E-08**	9.27E-02	7.18E-02	8.12E-02	**9.70E-09**	7.19E-02	7.46E-02

(a) Set tests

	Lead SNP	mtLMM-SV	stLMM-SV					
			basos	eos	lucs	lymphs	monos	neuts
3	*chrom8:131701894*	**1.03E-11**	2.24E-02	6.63E-01	2.99E-02	3.27E-02	**3.79E-11**	6.41E-01
12	*chrom9:902862*	**7.66E-14**	9.84E-03	5.96E-01	1.55E-04	**5.62E-14**	3.14E-01	8.13E-01
13	*chrom9:1177156*	**6.70E-14**	1.38E-02	5.48E-01	1.93E-04	**8.75E-14**	3.56E-01	9.23E-01

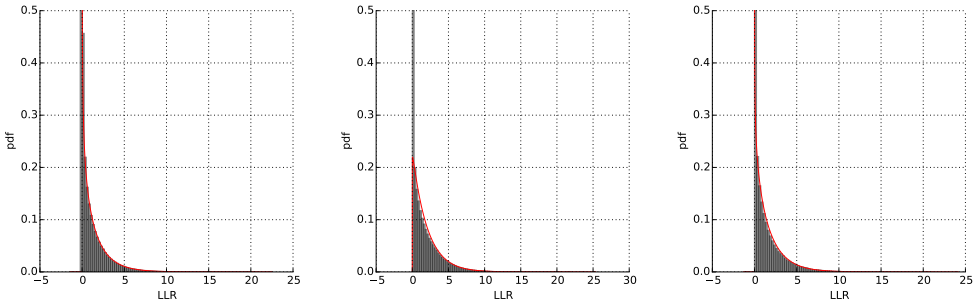
(b) Single variant tests

	start QTL	end QTL	length QTL
3	128089062	132222468	4.13Mb
12	865652	2619519	1.75Mb
13	865652	2619519	1.75Mb

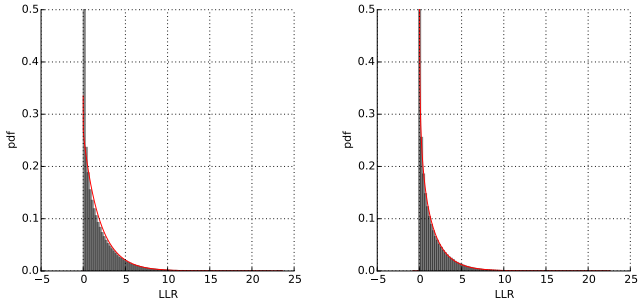
(c) QTL length in single-variant tests

Table B.11: **Tabular summary of QTLs identified by different set test and single-variant LMMs on the rat dataset.** The table shows significant associations ($\alpha < 0.01$, Bonferroni adjusted) found by different set test and single-variant LMMs in the analysis of the rat data. Only significant associations ($\alpha < 0.01$, Bonferroni adjusted) are in bold. (a) shows the results from all regions identified using set tests. (b) shows the lead variants from the regions identified by single-variant tests. (c) shows the length of QTL intervals identified by the multi-trait single-variant models, which were calculated as the genomic distance between the most extreme significant variants at the same locus.

B.2 Supplementary figures



(a) Repetition 1 **(b)** Repetition 2 **(c)** Repetition 3



(d) Repetition 4 **(e)** Repetition 5

Figure B.1: **Parametric Fit of the Null Distribution on the 1,000 Genomes dataset for mtSet.** The null distribution is fit by a mixture π of χ^2_0 and $a\chi^2_d$ test statistics using five genome-wide permutations. Although, we use only the top 10% of null test statistics for fitting the free parameters π, a, d, we found empirically that our fit works well for the complete range of the test statistics. Shown are the results for five different repetitions of four simulated phenotypes when only background effects are present.

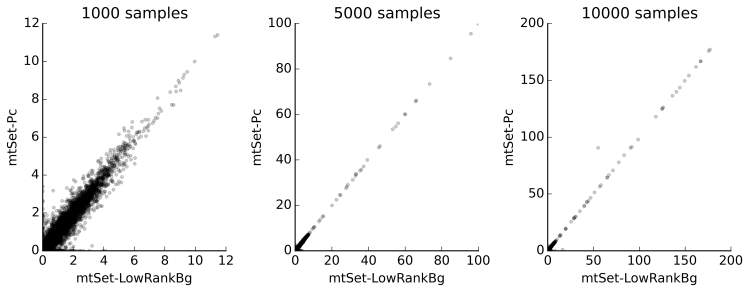
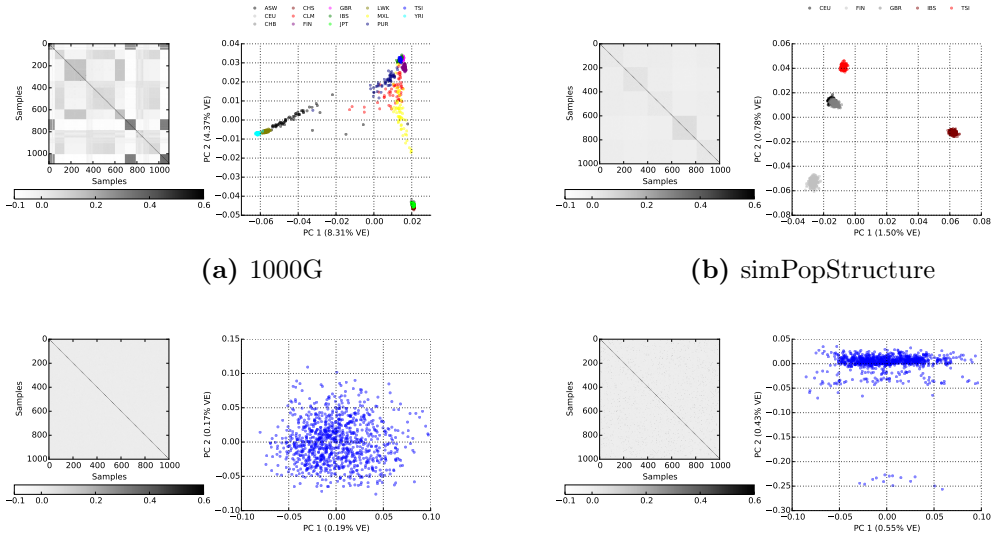


Figure B.2: **Comparison mtSet-PC and mtSet-LowRankBg.** Compared are likelihood ratio test statistics for mtSet-PC and mtSet-LowRankBG. For large cohorts, we find good concordance between both models. This shows that accounting for PCs as (REML) fixed effects or random effects yields similar results.



(a) 1000G **(b)** simPopStructure

(c) simUnrelated **(d)** simRelated

Figure B.3: **Characterization of the confounding structure in the four simulated datasets that were used to assess statistical calibration.** Shown are the empirical relatedness matrices as well as the scatter plots of the first two principal components. **(a)** Empirical genotype data of 1,000 individuals from 14 populations that are part of the 1000 genomes project (1000G). **(b-d)** Synthetic datasets based on genotypes from 1,000 Genomes Projects of European ancestry. In brief, each individual is assigned to n ancestors, randomly inheriting blocks of SNPs from its ancestors. By placing alternative restrictions on the ancestors (number of ancestors, ancestors are drawn from the same or different populations), datasets with different confounding population structure are obtained: **(b)** simPopStructure (kinship matrix is low-rank), **(c)** simUnrelated (kinship matrix is not structured) and **(d)** simRelated (kinship matrix is highly structure).

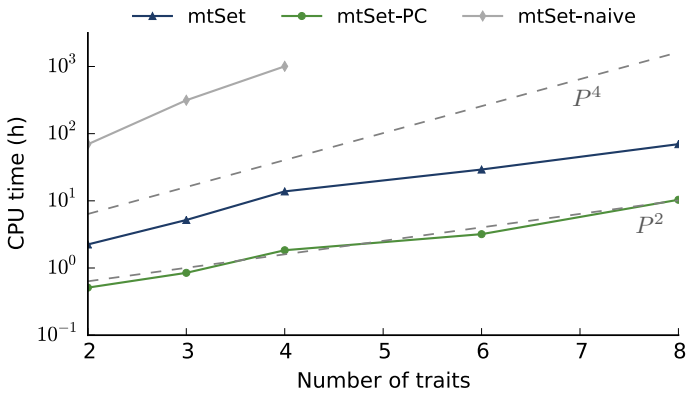


Figure B.4: **Computational cost for variable numbers of traits.** Shown is the extrapolated CPU time (h) to test associations on chromosome 20, considering a total of 3,975 windows (tests), on a simulated cohort with 1,000 individuals for increasing numbers of traits. Compared are mtSet and the approximate mtSet-PC model. Naive denotes the runtime for a standard LMM package, which scales cubical in the number of traits times the number of individuals. Runtime estimates were obtained from a single core of an Intel Xeon CPU E5-2670 2.60 GHz processor.

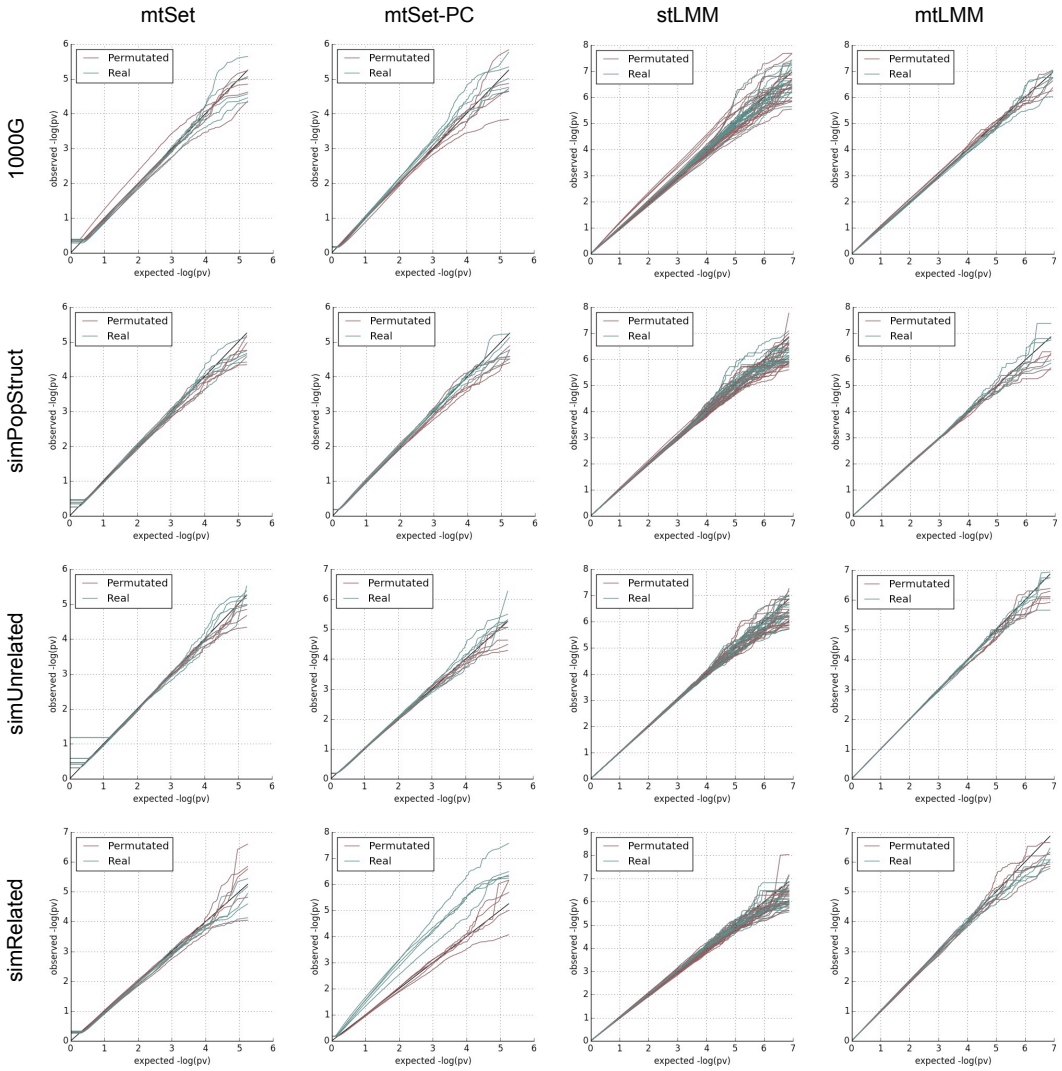
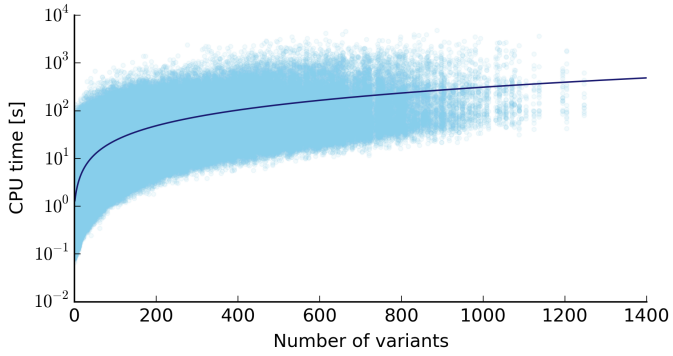
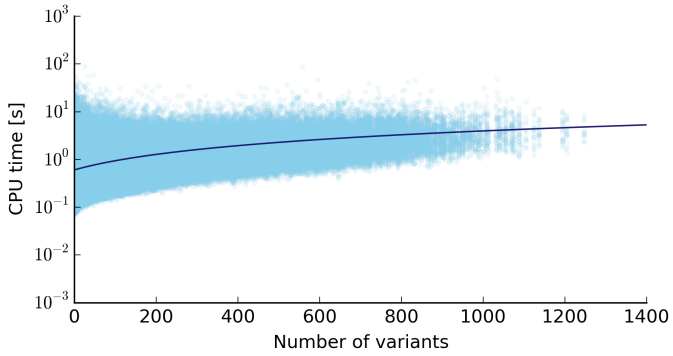


Figure B.5: **Statistical calibration of mtSet, mtSet-PC, stLMM-SV and mtLMM-SV on simulated datasets with different confounding structure.** Shown are QQ-plots for simulated data when only background effects (no causal variants) were simulated when considering alternative degrees of population structure and relatedness (see **Fig. ??**). Compared are a single trait singe SNP LMM (stLMM-SV), a multi-trait single-SNP LMM (mtLMM-SV) as well as mtSet and the PC-based approximation (mtSet-PC).
From left to right: mtSet, mtSet-PC, stLMM-SV and mtLMM-SV.
From top to bottom: 1000 Genomes (real genotypes), simPopStructure, simUnrelated, simRelated. Whereas the calibration of mtSet, stLMM-SV and mtLMM-SV were not affected by the type of confounding, mtSet-PC was not able to account for complex relatedness structures (bottom row, right-most plot).



(a) mtSet



(b) mtSetPC

Figure B.6: **Scalability of mtSet as a function of the number of variants in the set component.** Shown is computational time to fit a single window using mtSet (**a**) and mtSet-PC (**b**) (randomly drawn from chrom 20, 1,000 Genomes dataset), considering a *Intel(R) Xeon(R) CPU E5-2670 0 @ 2.60GHz*, for windows with increasing numbers of variants. Runtimes are reported for windows of varying size (1kb-200kb) using simulated data generated using the default parameter settings (see also Table B.4).

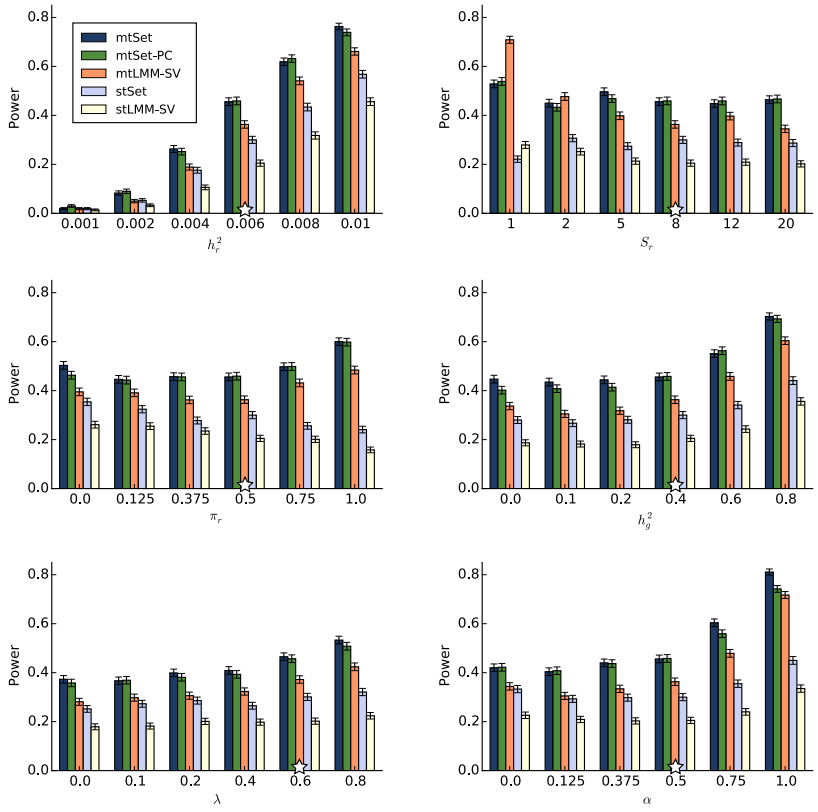
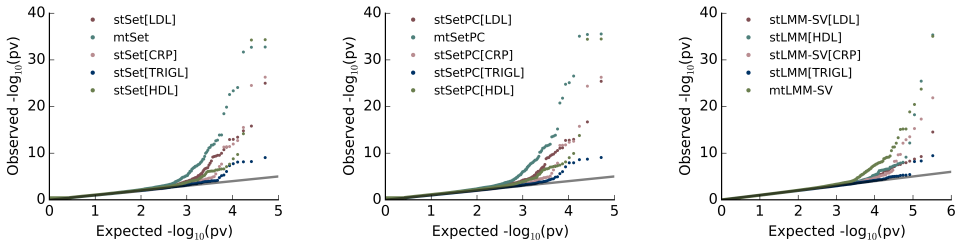


Figure B.7: **Power comparison of alternative methods on simulated data from 1000 genomes project genotypes.** Shown is power at 10% family-wise error rate for mtSet, stSet, mtSet-PC, mtLMM-SV and stLMM-SV varying different simulation parameters. Specifically, we altered the proportions of variance explained by the region (h_r^2), the numbers of causal variants in the region (S_r), the percentages of shared causal variants (π_r), the proportions of variance explained by genetic background (h_g^2), the percentage of residual variance explained by hidden confounders (λ), and the percentage of background and residual signal that is shared across traits (α) (see also Table B.4).



(a) **(b)** **(c)**

Figure B.9: **QQ-plots for blood lipid levels on the NFBC dataset.** All methods show good calibration. Genomic control is λ(mtLMM-SV) = 0.979, λ(stLMM-SV[CRP]) = 0.995, λ(stLMM-SV[LDL]) = 0.996, λ(stLMM-SV[HDL]) = 1.001 and λ(stLMM-SV[TRIGL]) = 0.978 for the single-variant methods, λ(mtSet) = 1.001 and λ(mtSetPC) = 0.989 for the set methods.

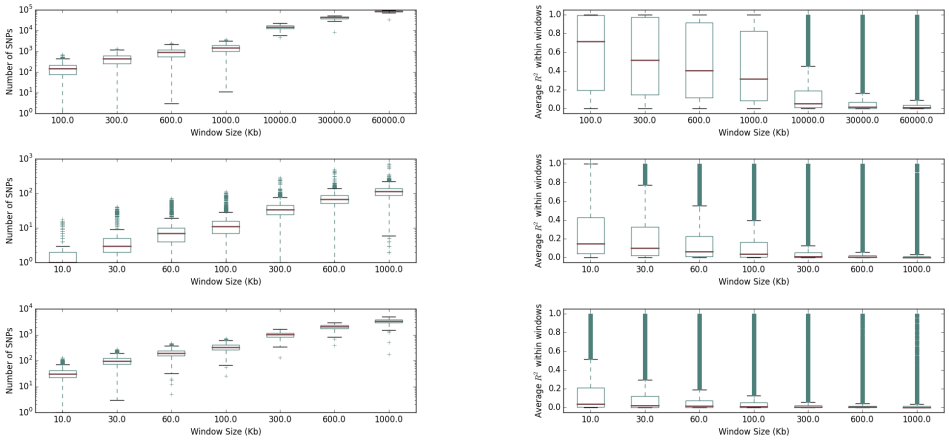
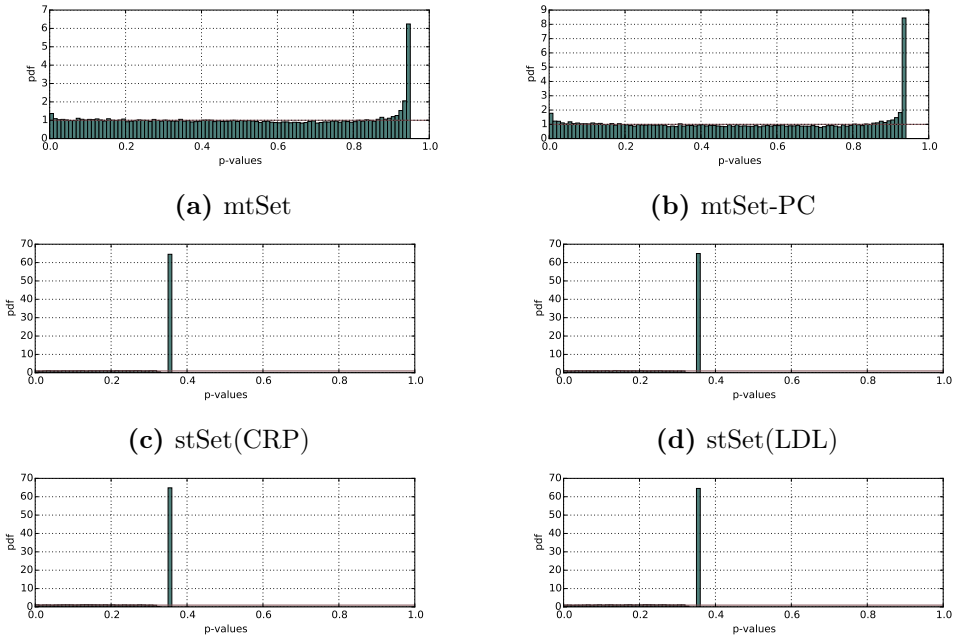


Figure B.8: **Distribution of the number of variants within testing regions as well as the squared intra-SNP correlation coefficient, both as a function of the considered window size. Left column:** Dependency between window sizes and number of SNPs. **Right column:** Dependency between window sizes and SNP-SNP squared correlation within windows. **From top to bottom:** Rat datasets, NFBC data, 10000 Genomes (chromosome 20). The computational cost of mtSet depends on the number of (unique) SNPs in a window. In the experiments, we considered 100kb windows for NFBC, 1mb windows for rat and 30kb windows for the 1,000 genomes data.

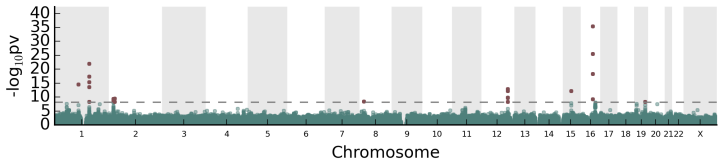


(a) mtSet **(b)** mtSet-PC

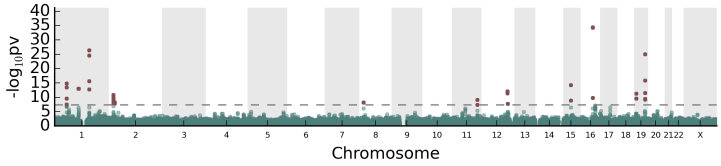
(c) stSet(CRP) **(d)** stSet(LDL)

(e) stSet(HDL) **(f)** stSet(TRIGL)

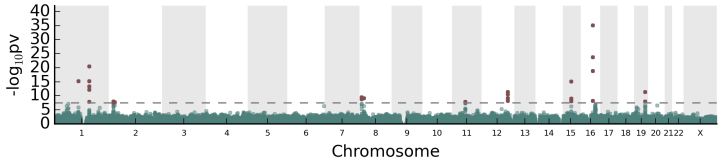
Figure B.10: **Histogram of p-values obtained from set tests applied to four blood lipid levels on the NFBC dataset. Top row:** multi-trait set tests (mtSet, mtSet-PC) applied to all four traits jointly.
Bottom two rows: single-trait set test (stSet) applies to individual traits. The peaks in the histograms is common to all set tests and is a result of the constrained optimization of the marginal likelihood: in these instances the set component variance parameter is zero, the bound of the optimization. The location of the peak is a function of the mixture coefficients of the parametric null distribution fit.



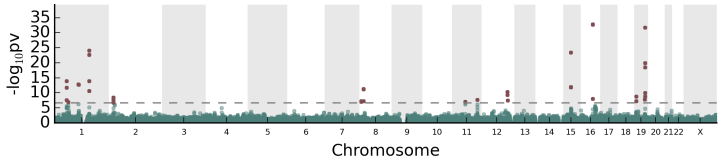
(a) stLMM-SV: minimum p-value



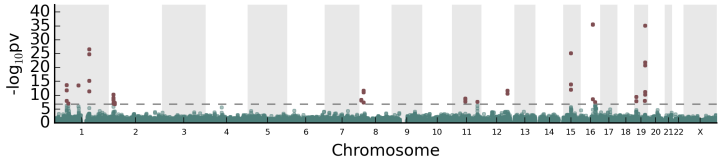
(b) stSet: minimum p-value



(c) mtLMM-SV

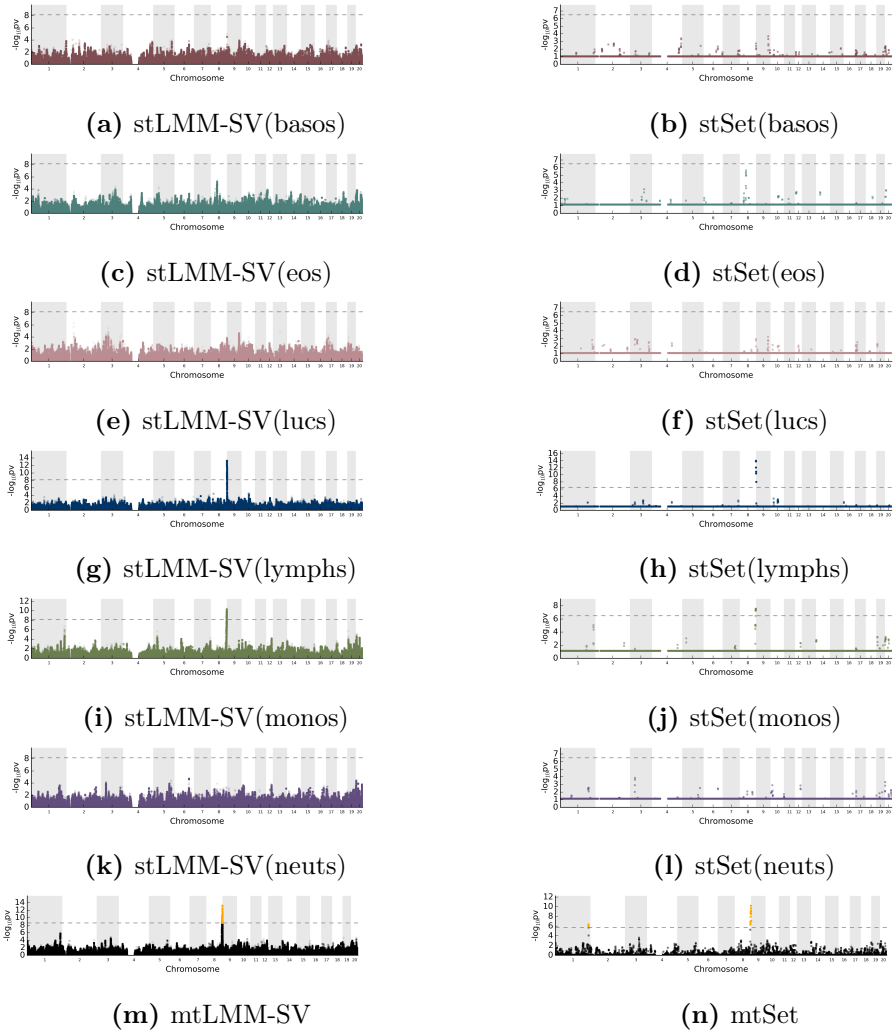


(d) mtSet



(e) mtSet-PC

Figure B.11: **Manhattan plots for blood lipid levels on the NFBC dataset.** We report the minimal p-values over all phenotypes (CRP, LDL, HDL and TRIGL) for the single-trait methods stLMM-SV (**a**) and stSet (**b**). The manhattan plots for mtLMM-SV, mtSet and mtSet-PC are shown in (**c**,**d**) and (**e**). The model mtSet-PC recovers all associations that are found by its competitors (stLMM-SV,mtLMM-SV and stSet) and two additional associations: the first one is on chromosome 1 (shared with mtSet) and the second one on chromosome 16 .



(a) stLMM-SV(basos) **(b)** stSet(basos)

(c) stLMM-SV(eos) **(d)** stSet(eos)

(e) stLMM-SV(lucs) **(f)** stSet(lucs)

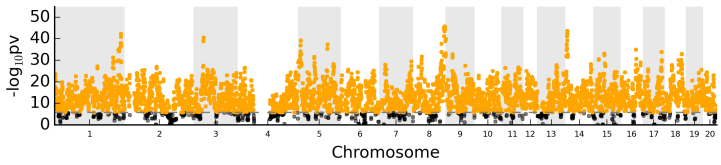
(g) stLMM-SV(lymphs) **(h)** stSet(lymphs)

(i) stLMM-SV(monos) **(j)** stSet(monos)

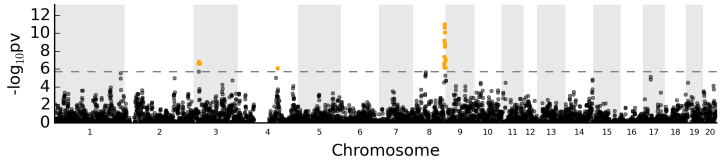
(k) stLMM-SV(neuts) **(l)** stSet(neuts)

(m) mtLMM-SV **(n)** mtSet

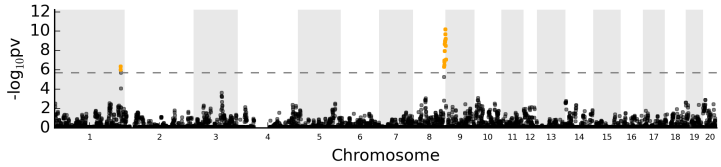
Figure B.12: **Manhattan plots for quantitative traits related to basal haematology on rat.** (**a**, **c**, **e**, **g**, **i**, **k**) show manhattan plots for basophils (basos), eosinophils (eos), large unstained cells (luc), lymphocytes (lymphs), monocytes (monos) and neutrophils (neuts) obtained using stSet-SV while (**b**, **d**, **f**, **h**, **j**, **l**) show manhattan plots for the same traits from stSet. Finally, (**m**) and (**n**) show manhattan plots for multi-trait mixed model (mtLMM-SV) and mtSet respectively. Notice that the horizontal line in manhattan plots for stSet contains all regions for which the log-likelihood ratio is ≈ 0 and it is a consequence of the LLR test statistics, which follows approximately a mixture of chi squared.



(a) Multi-Trait Set Test - without the relatedness component (mtSet-noBg).



(b) Multi-Trait Set Test - top 30 principal components (mtSet-PC).



(c) Multi-Trait Set Test - including the relatedness component (mtSet).

Figure B.13: **Manhattan plots for alternative set tests applied to six phenotypes related to basal haematology on the rat dataset.** **a**, manhattan plot obtained when omitting the relatedness component (mtSet-noBg). **b**, equivalent manhattan plot when using mtSet-PC to correct for population structure. **c**, corresponding results when including the relatedness component (mtSet). Only the full model that includes the relatedness component (**c**) is able to comprehensively correct for relatedness.

C | Supplementary results for: Testing for polygenic interactions using set tests

C.1 Supplementary tables

Method	Paper	LMM type	Multiple variants	Fully observed designs	Stratified samples	Signal heterogeneity across contexts	Variance decomposition	Variant type
iSet	-	multivariate	✓	✓	✓	✓	✓	common
MTMM	*Korte et al (2012)*	multivariate	✗	✓	✗	✓	✗	common
GESAT / iSKAT	*Lin et al (2013)* *Lin et al (2016)*	univariate	✓	✗	✓	✗	✗	common / rare
SimReg	*Tzeng et al (2011)* *Zhao et al (2015)*	univariate	✓	✗	✓	✗	✗	common / rare
Turkey's 1dof test	*Chatterjee et al (2006)*	univariate	✓	✗	✓	✗	✗	common

Table C.1: **Comparison table of iSet and related models for interaction testing.**

Effect	Function of parameters	Variance explained
Region effect	v_r	2%
Shared relatedness effects	αv_{bg}	24%
Independent relatedness effects	$(1-\alpha)v_{bg}$	16%
Shared effects from hidden factors	$\alpha\beta(1-v_r-v_{bg})$	17.4%
Independent effects from hidden factors	$(1-\alpha)\beta(1-v_r-v_{bg})$	11.6%
IID noise	$(1-\beta)(1-v_r-v_{bg})$	29%

(a) Variance explained by the different contributions

Parameter										
Number of causal variants	1	2	**4**	8	12	20	-	-	-	-
Proportionality factor (extent of rescaling-GxC)	-0.5	-0.3	-0.2	-0.1	0	0.1	**0.2**	0.3	0.5	0.7

(b) Parameter values for rescaling-GxC simulations

Parameter										
Number of causal variants (across contexts)	2	**4**	8	12	20	-	-	-	-	-
Region effect correlation (extent of heter-GxC)	-1, -.8	-.8, -.6	-.6, -.4	-.4, -0.2	-.2, 0	0, .2	.2, .4	**.4, .6**	6, .8	.8, 1

(c) Parameter values for general-GxC simulations

Table C.2: **Simulation settings.** Simulated Phenotype data were generated as sum of an effect from the variants of the genetic region, an effect from relatedness / population structure, an effect from $K = 10$ unmeasured hidden confounders and iid noise. We fixed the variance explained by the region ($v_r = 2\%$), the fraction of shared background signal ($\alpha = 0.6$), the fraction of residual variance that is explained by the the hidden factors (β=0.5) while varying the number of causal variants and the extents of rescaling-GxC and heterogeneity-GxC. **(a)** The contributions to phenotypic variance of all simulated effects. **(b,c)** The values of parameters considered respectively for simulations of rescaling-GxC and heterogeneity-GxC effects. Each of the parameters in **(b,c)** was varied while keeping others at the default value (bold face).

	Opposite Direction	Same Direction	Fold Change	P value
heter / No heter	3 / 8 (37.5%)	96 / 1327 (7.2%)	5.2	$3.51\ 10^{-2}$

(a) IFN

	Opposite Direction	Same Direction	Fold Change	P value
heter / No heter	2 / 21 (9.5%)	86 / 1002 (8.5%)	1.11	0.56

(b) LPS2

	Opposite Direction	Same Direction	Fold Change	P value
heter / No heter	6 / 13 (46.2%)	79 / 1239 (6.4%)	7.2	$7.94\ 10^{-4}$

(c) LPS24

	Opposite Direction	Same Direction	Fold Change	P value
heter / No heter	11 / 42 (26.1%)	261 / 3560 (7.3%)	3.6	$8.9\ 10^{-4}$

(d) All

Table C.3: **Enrichment of heterogeneity-QTLs in opposite direction QTLs** Breakdown of probe / stimulus pairs with shared lead variants, stratified by concordance of the effect direction (opposite-direction versus same-direction eQTLs) and significance of the heterogeneity-GxC test (heter vs No heter) in naïve/IFN (a), naïve/LPS2 (b), naive/LPS24 (c) and aggregating across all stimuli (d).

Trait	Chrom	Start	End	Type	Pv stSet	Pv mtSet	Pv iSet	Pv GESAT	minPv mtLMM-int
crp3dec	1	40400000	40500000	Gene-by-gender	1.89E-03	1.42E-07	1.47E-06	5.98E-05	1.25E-05
FS_KOL_L	3	121800000	121900000	Gene-by-gender	3.66E-01	1.72E-05	3.74E-06	4.82E-06	3.18E-05
FS_TRIGL	17	4500000	4600000	Association	1.89E-06	4.73E-08	3.53E-04	5.24E-03	1.80E-03
FS_KOL_H	11	47250000	47350000	Association	9.94E-07	3.31E-07	8.56E-03	1.90E-02	5.03E-03

Table C.4: **Tabular summary of interaction and association loci in blood lipid profiles from NFBC1966.** Tabular summary of the interaction and association loci discussed in the Section 4.4.3.

C.2 Supplementary figures

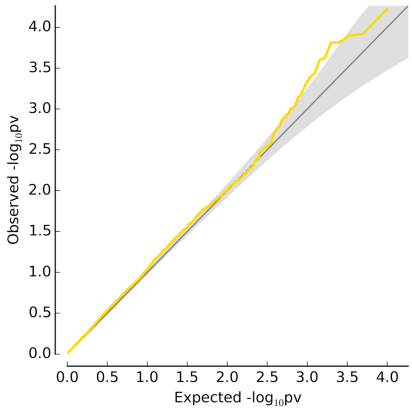
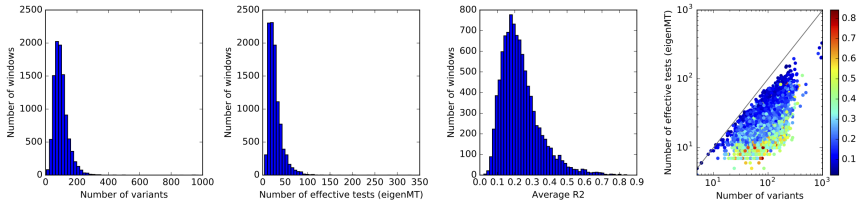
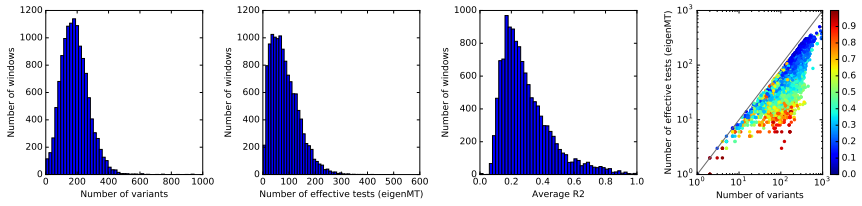


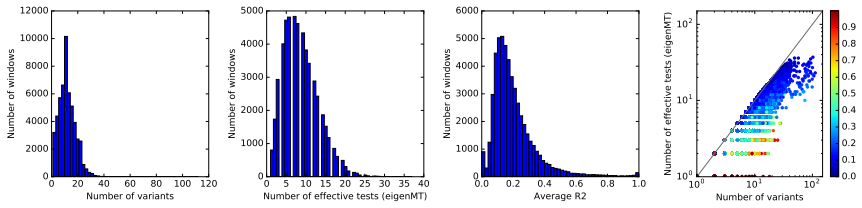
Figure C.1: **Statistical calibration of iSet-het when only rescaling-GxC effects are simulated.** Shown is the QQ plot for the P values obtained from the heterogeneity-GxC interaction test (iSet-het) when simulating rescaling-GxC (without heterogeneity-GxC). P values are pooled across all simulations in **Fig. 4.4c** (where exclusively rescaling-GxC effects were simulated).



(a) 1000G



(b) Stimulus eQTL dataset



(c) NFBC

Figure C.2: **Distribution of the number of variants, number of effective tests estimated by eigenMT and average squared correlation within the testing regions in the different datasets.** From left to right: distribution of the number of variants across the analysed regions; distribution of the number of effective tests as estimated by eigenMT; distribution of the average pair-wise squared Pearson correlation (R^2) across all variants in each region; scatter plot of the number of effective tests versus the number of variants. Shown in colour is the within-region average correlation (R^2) across all pairs of variants. From top to bottom: (**a**) 10,000 30kb regions from the simulated data (1,000 individuals) based on 1000 Genomes individuals, (**b**) 100kb cis regions (centred on the TSS) considered in the cis stimulus eQTL analysis (288 individuals) and (**c**) 100kb regions considered in genotype-sex interaction analysis in NFBC (5,402 individuals).

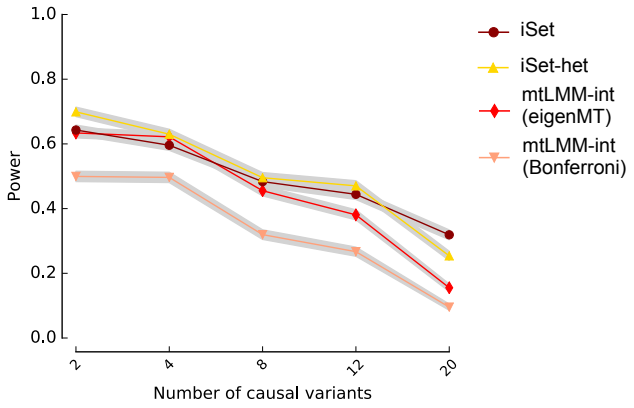


Figure C.3: **Power of iSet and iSet-het when simulating heterogeneity-GxC effects and increasing numbers of causal variants.** Shown is the power of iSet and a single-variant interaction test (mtLMM-int) to detect GxC interactions when simulating heterogeneity-GxC and increasing the number of causal variants. We also report power of iSet-het to detect heterogeneity-GxC.

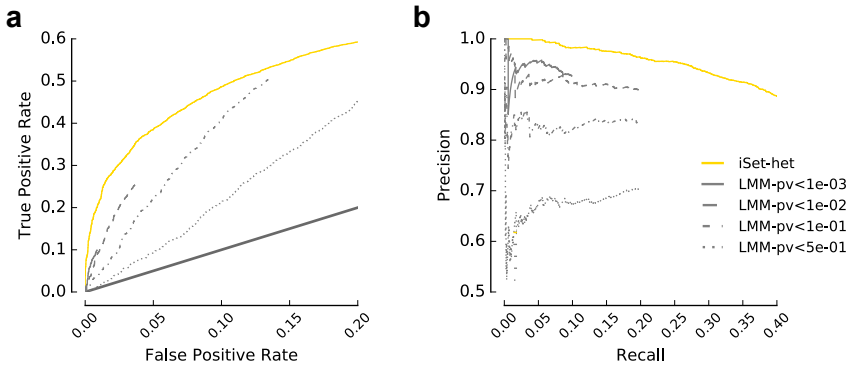
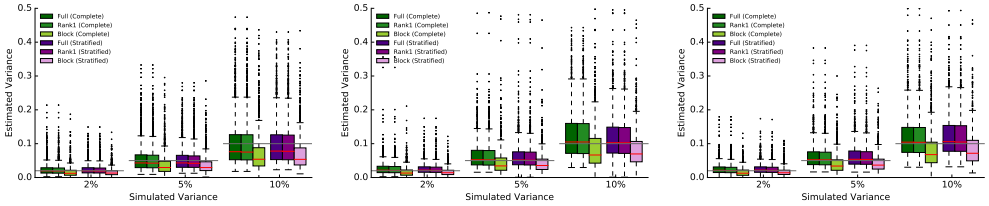
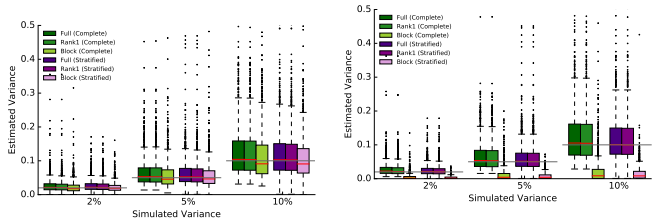


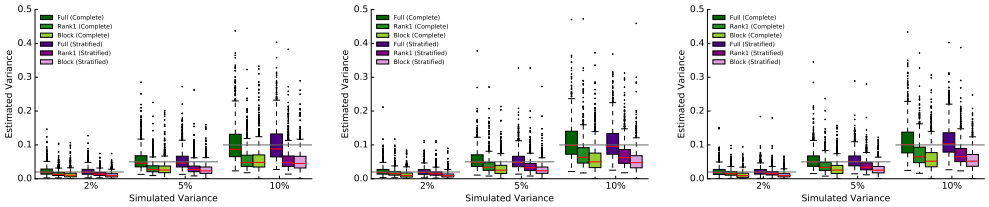
Figure C.4: **Comparison of iSet-het and single-variant strategies for discriminating rescaling from heterogeneity-GxC.** Receiver operating curve **(a)** and precision-recall curve **(b)** when using different approaches to discriminate heterogeneity-GxC. Considered was the iSet test to score the extent of heterogeneity (iSet-het) and a baseline approach based on single-trait single-variant LMMs. Briefly, the considered score is $1 - r^2$ (where r is the Pearson correlation between lead variants identified in each context) for regions with significant associations in both contexts (P-value thresholds 0.5, 0.01, 1e-3, 1e-4). Regions that were not marginally significant in either one of the two contexts were assigned a score of zero. In a ROC curve the true positive rate is plotted against the false positive rate threshold. In a PR curve precision of the model (which is the fraction of retrieved cases that are positive) is plotted against recall (the fraction of positive cases that are retrivied).



(a) Resc-GxC (0.2)-1 causal **(b)** Resc-GxC (0.2)-4 causal **(c)** Resc-GxC (0.2)-8 causal



(d) Resc-GxC (0.5)-4 causal **(e)** Resc-GxC (-0.5)-4 causal



(f) Het-GxC-2 causal **(g)** Het-GxC-4 causal **(h)** Het-GxC-8 causal

Figure C.5: **Assessment of genetic variance estimates from iSet using different covariance models.** Shown are the estimates of the genetic variance explained by the set component across different simulated settings when considering alternative covariance models. In particular, shown are the variance component estimates when alternatively considering a full-rank covariance (full, general case), a rank-one covariance (rank1, only rescaling-GxE) and a block covariance matrix (block, which models only persistent genetic effects). Both designs with fully observed cohorts (complete - 1,000 individuals and 2 contexts for a total of 2,000 trait measurements) and stratified samples (stratified - 2,000 individuals and 2 contexts for a total of 2,000 trait measurements) are considered. Only the full-rank iSet model yields variance component estimates that are calibrated across all simulated scenarios. In particular, we considered scenarios with either rescaling-GxC effects (where we varied the number of causal SNPs and the proportionality factor of the effect sizes across the two contexts) or heterogeneity-GxC (where we vary the number of SNPs). For each simulated scenario we considered 1,000 simulated regions and altered the variance explained by the region (we consider the values 2%, 5% and 10%). Grey horizontal lines denote the true simulated local genetic variance.

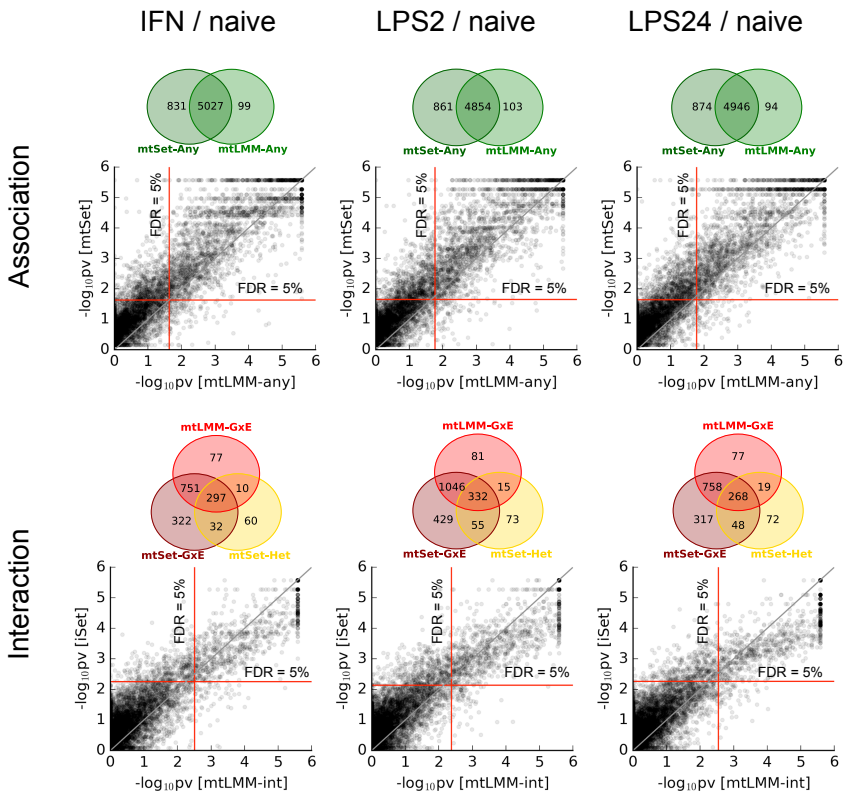


Figure C.6: **Comparison of single-variant methods and set tests on the monocyte stimulus eQTL data.** Shown are the scatter plots of the -$\log_{10}$P values from single-variant LMMs and set tests for association tests (mtLMM vs mtSet) and interaction tests (mtLMM-int vs iSet) for different stimulus contexts (IFN / naive, LPS-2h / naive, LPS-24h / naive). Region-based P values for single-variant models are minimum P-value across variants in the region, adjusted for the effective number of tests (estimated using eigenMT). Venn diagrams on the top of individual panels show the overlap of genes with significant associations or interactions identified using alternative methods (5% FDR).

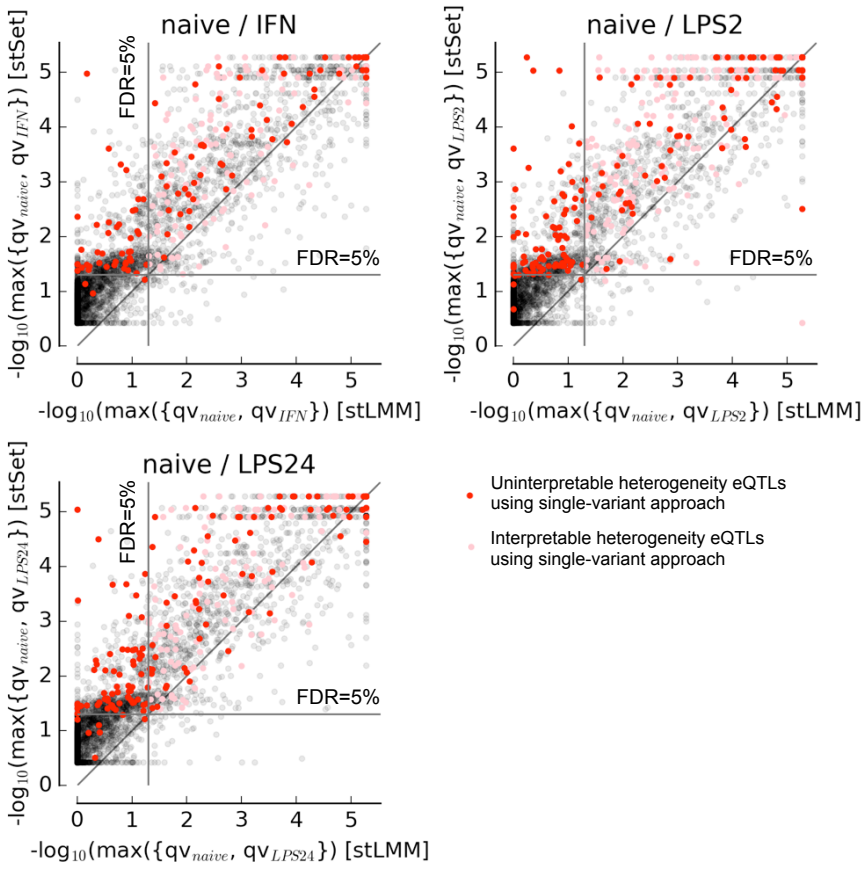


Figure C.7: **Comparison of adjusted P values obtained using either a univariate set set or a univariate single-variant test.** Scatter plot of the adjusted P values across pairs of contexts (independent analysis), comparing set tests and single-variant tests for different stimulus pairs. Heterogeneity-GxC cases without clear single-variant interpretation (highlighted in red) tend to be more significant when using set tests, suggesting that differences in power hamper the single-variant annotation of these eQTLs.

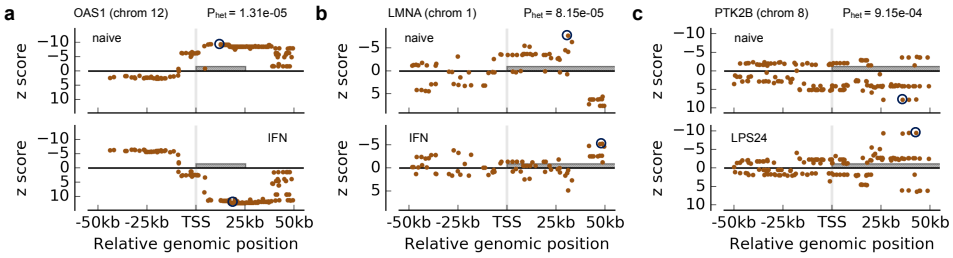
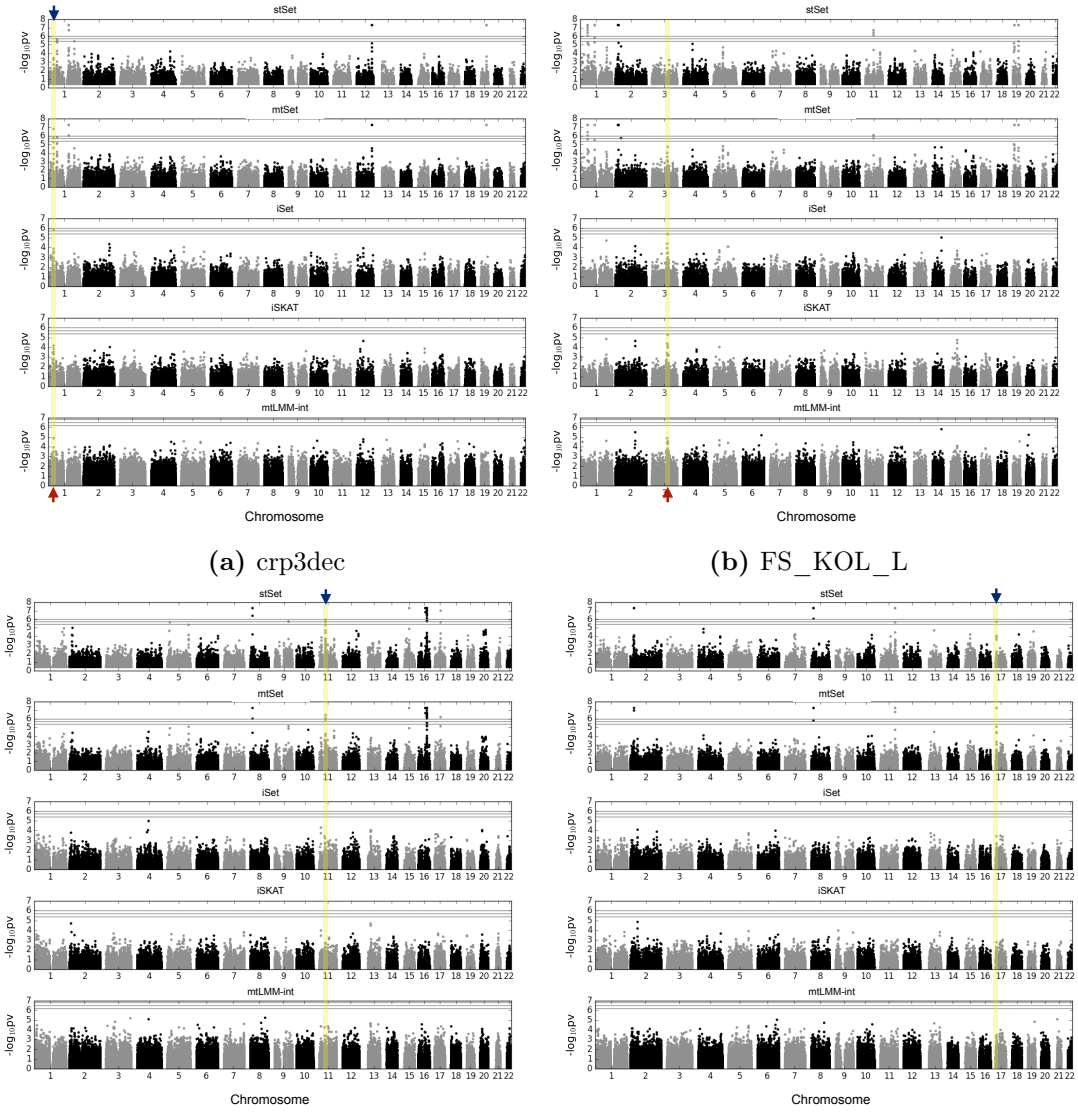


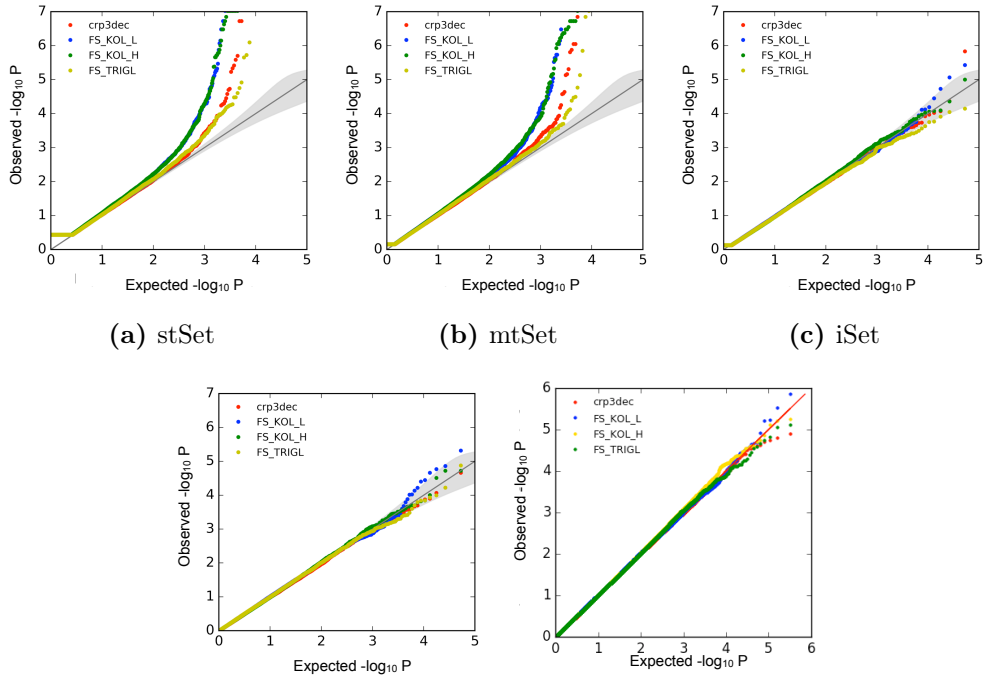
Figure C.8: **Examples of opposite-effect eQTLs with significant heterogeneity-GxC effects.** Shown is the z-score statistics at *cis* variants for OAS (**a**), LMNA (**b**) and PTK2B (**b**) across the contexts showing opposite effects. While the three examples are identified as opposite effects when using single-variant methods (Fairfax et al., 2014), iSet identifies significant heterogeneity-GxC, indicating changes in the configuration of causal variants. Lead variants in individual contexts are annotated using circles and are in high LD ($r^2 > 0.8$).



(a) crp3dec **(b)** FS_KOL_L

(c) FS_KOL_H **(d)** FS_TRIG

Figure C.9: **Manhattan plots when applying alternative methods to lipid levels in NFBC1966.** Shown are Manhattan plots for C-reactive protein (crp3dec, a), LDL cholesterol (FS_KOL_L, b), HDL cholesterol (FS_KOL_H, c), and triglycerides (FS_TRIG, d) obtained from univariate set tests ignoring sex-specific differences (stSet), an association test that accounts for differences in genetic effect across strata (mtSet), iSet, GESAT and single-variant interaction test (mtLMM-int). Red arrows indicate the interaction effects that are discussed in the Section 4.4.3. Blue arrows indicated associations that can only be detected when modelling differences in effect sizes across strata (mtSet vs stSet).



(a) stSet **(b)** mtSet **(c)** iSet

(d) GESAT **(e)** mtLMM-int

Figure C.10: **QQ plots when applying alternative methods to lipid levels in NFBC1966.** Shown are the QQ plots for C-reactive protein (crp3dec), LDL cholesterol (FS_KOL_L), HDL cholesterol (FS_KOL_H), and triglycerides (FS_TRIG) obtained using a univariate association set tests ignoring sex (stSet, a), an association test modelling sex-specific genetic effects (mtSet, b), iSet (c), GESAT (d) and single-variant interaction test (mtLMM-int).

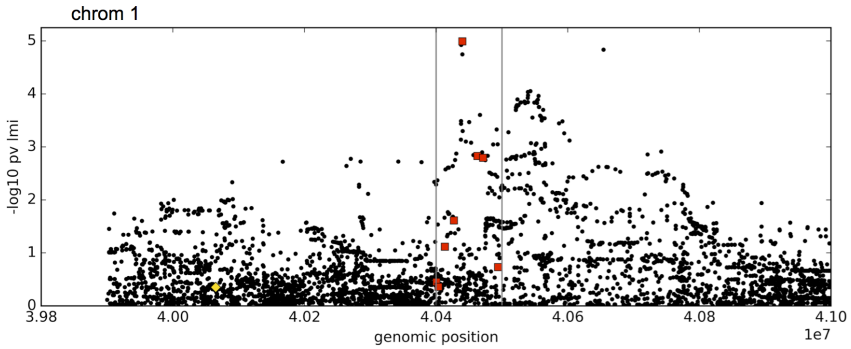


Figure C.11: **Manhattan plot in the interaction locus for C-reactive protein using single-variant interaction tests on imputed variants.** Shown is the Manhattan plot for C-reactive protein obtained from single-variant interaction tests applied to common variants (MAF> 0.5%) on imputed data. The vertical grey lines indicate the 100kb region with significant genotype-sex interaction (FWER=10%) when using iSet. Non-imputed typed variants are highlighted in red, showing that for this locus imputation strategies did not increase the power of single-variant methods.

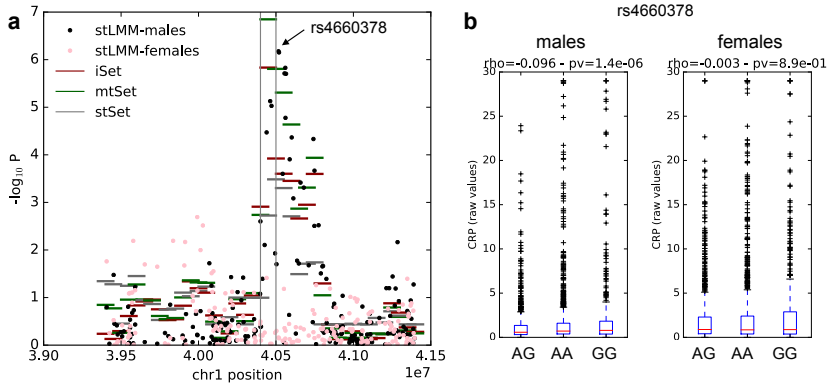


Figure C.12: **The interaction for C-Reactive protein on chromosome 1 is a male-specific effect.** (**a**) Local Manhattan plot (1Mb around significant region) for single-variant association tests, either considering males (black) or females (pink). For comparison, we also show P-values from the iSet (red), mtSet (green) and stSet (grey). (**b**) C-Reactive protein level stratified by different alleles of rs4660378 (lead SNP identified in the analysis using male individuals only). rho corresponds to the Spearman correlation coefficient and the corresponding P value of the correlation test, both for male and female individuals.

D | Supplementary material for: Flexible LINear MIXed models

D.1 Covariance functions and Gaussian processes

As we have seen in Chapter 5, the LIMIX inference framework builds on the concept of Gaussian process and covariance function. As described below, the Gaussian process can be seen as an extension of the linear mixed model (LMM). While I here introduce the Gaussian process and the covariance function in an intuitive manner, for an in-depth discussion I refer the interested reader to Rasmussen (2006).

Covariance function. Let $\mathbb{F}$ denote the input feature space. A function $\kappa : (x, x') \in \mathbb{F}^2 \to \mathbb{R}$ is a covariance function if and only if $\forall N$ and $\forall \boldsymbol{x} \in \mathbb{F}^N$ the matrix $\boldsymbol{K}_{\boldsymbol{x}} = [\kappa(x_i, x_j)] \in \mathbb{R}^{N \times N}$ is positive-semidefinte. Covariance functions are also known as kernels.

Gaussian process. A Gaussian process (GP) is a distribution over functions characterised by a mean function and a covariance function. Let $\mathbb{F}$ denote the input feature space. A function $f : \mathbb{F} \to \mathbb{R}$ follows a GP distribution with mean function $m : \mathbb{F} \to \mathbb{R}$ and covariance function $\kappa : (x, x') \in \mathbb{F}^2 \to \mathbb{R}$ if and only if $f(\boldsymbol{x}) \sim \mathcal{N}\left(m(\boldsymbol{x}), \boldsymbol{K}_{\boldsymbol{x}}\right)$ $\forall N$ and $\forall \boldsymbol{x} \in \mathbb{F}^N$, where $f(\boldsymbol{x}) = [f(\boldsymbol{x}_i)] \in \mathbb{R}^N$, $m(\boldsymbol{x}) = [m(\boldsymbol{x}_i)] \in \mathbb{R}^N$ and $\boldsymbol{K}_{\boldsymbol{x}} = [\kappa(x_i, x_j)] \in \mathbb{R}^{N \times N}$. Intuitively, the Gaussian process, its mean function and its covariance function are infinite-dimensional generalisation of a multivariate Gaussian distribution, its mean vector and its covariance matrix, respectively.

Relationship with LMMs. Let N denote the number of samples, Q the number of input variables, $\boldsymbol{y} \in \mathbb{R}^N$ the outcome vector and $\boldsymbol{Z} \in \mathbb{R}^{N \times Q}$ the design matrix of the input variables. In this example, the input feature space is $\mathbb{F} = \mathbb{R}^Q$. Let us introduce

a function of the input feature $\phi_{\boldsymbol{\theta}} : \mathbb{R}^{Q} \in\rightarrow \mathbb{R}^{\mathcal{Q}}$ that depends on some parameters $\boldsymbol{\theta}$. Notice that neither the explicit form of $\phi_{\boldsymbol{\theta}}$ nor the dimension of $\mathcal{Q}$ are known. Denoting with $\boldsymbol{z}_q$ the q-th row of $\boldsymbol{Z}$, Let us consider a LMM where the outcome is linear in the transformed input variables

$$\boldsymbol{y} = \begin{bmatrix} \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_1\right)^\top \\ \vdots \\ \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_N\right)^\top \end{bmatrix} \boldsymbol{b} + \boldsymbol{\psi}, \qquad \boldsymbol{\psi} \sim \mathcal{N}\left(\mathbf{0}\,,\, \sigma_e^2 \boldsymbol{I}_N\right), \tag{D.1}$$

where $\boldsymbol{b} \in \mathbb{R}^{\mathcal{Q}}$ is modelled as random effect, $\boldsymbol{b} \overset{\text{iid}}{\sim} \mathcal{N}\left(0\,,\,1\right)$. Marginalising $\boldsymbol{b}$ out, we obtain

$$\boldsymbol{y} \sim \mathcal{N}\left(\mathbf{0}\,,\, \begin{bmatrix} \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_1\right)^\top \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_1\right) & \dots & \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_1\right)^\top \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_N\right) \\ \vdots & \ddots & \vdots \\ \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_N\right)^\top \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_1\right) & \dots & \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_N\right)^\top \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_N\right) \end{bmatrix} + \sigma_e^2 \boldsymbol{I}_N \right). \tag{D.2}$$

The last equation shows that the non-linear relationship between the output variables and the original input variables only depends on the the dot product in the transformed space. This dot product can be defined using a covariance function $\kappa_{\boldsymbol{\theta}}$ as follows

$$\kappa_{\boldsymbol{\theta}}(\boldsymbol{z}_{i,:}, \boldsymbol{z}_{j,:}) = \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_{i,:}\right)^\top \phi_{\boldsymbol{\theta}}\left(\boldsymbol{z}_{j,:}\right), \tag{D.3}$$

where $\boldsymbol{\theta}$ are the parameters of the covariance function. In sum, covariance functions can be used to define non-linear relationships between the output and the input variables (Schölkopf and Smola, 2002).

D.2 Basic covariance models

Fixed-form covariance When the covariance structure between samples is known, only a positive scale parameter e^a needs to be inferred. The covariance function and its derivative are respectively

$$\boldsymbol{K}_{\boldsymbol{\theta}} = e^a \boldsymbol{K}_0 \tag{D.4}$$

$$\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial a} = e^a \boldsymbol{K}_0, \tag{D.5}$$

where $\boldsymbol{\theta} = \{a\}$ and $\boldsymbol{K}_0$ is a fixed covariance matrix.

Freeform covariance A freeform covariance is the most general form of covariance matrix. To represent a general covariance and ensure that it is positive semidefinite, we use the Cholesky parametrisation. Briefly, indicating with d the dimension of the covariance and with $\theta \in \boldsymbol{R}^{\frac{1}{2}d(d+1)}$ the non-zero entries of a $d \times d$ lower triangular matrix $\boldsymbol{L}_{\boldsymbol{\theta}}$, the freeform matrix and its gradients are

$$\begin{aligned} \boldsymbol{K}_{\boldsymbol{\theta}} &= \boldsymbol{L}_{\boldsymbol{\theta}} \boldsymbol{L}_{\boldsymbol{\theta}}^{\top} && \text{(D.6)} \\ \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_j} &= \frac{\partial \boldsymbol{L}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_j} \boldsymbol{L}_{\boldsymbol{\theta}}^{\top} + \boldsymbol{L}_{\boldsymbol{\theta}} \frac{\partial \boldsymbol{L}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_j}^{\top} && \text{(D.7)} \end{aligned}$$

Lowrank covariance A lowrank covariance is a general covariance matrix having rank $r < d$. Indicating with $\theta \in \boldsymbol{R}^{rd}$ the elements of a $d \times r$ matrix $\boldsymbol{X}_{\boldsymbol{\theta}}$, the lowrank matrix and its gradients are

$$\begin{aligned} \boldsymbol{K}_{\boldsymbol{\theta}} &= \boldsymbol{X}_{\boldsymbol{\theta}} \boldsymbol{X}_{\boldsymbol{\theta}}^{\top} && \text{(D.8)} \\ \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_j} &= \frac{\partial \boldsymbol{X}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_j} \boldsymbol{X}_{\boldsymbol{\theta}}^{\top} + \boldsymbol{X}_{\boldsymbol{\theta}} \frac{\partial \boldsymbol{X}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_j}^{\top} && \text{(D.9)} \end{aligned}$$

Diagonal covariance A diagonal covariance matrix and its gradients are

$$\begin{aligned} \boldsymbol{K}_{\boldsymbol{\theta}} &= \operatorname{diag}\left(e^{a_0}, \ldots, e^{a_d}\right) && \text{(D.10)} \\ \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial a_j} &= \operatorname{diag}\left(0, \ldots, \overbrace{e^{a_j}}^{j\text{-th element}}, \ldots, 0\right), && \text{(D.11)} \end{aligned}$$

where $\boldsymbol{\theta} = \{a_0, \ldots, a_d\}$.

Covariance functions Indicating with N the number of individuals, let $\boldsymbol{X}$ be the $N \times D$ input matrix for d features. As we have seen in Section D.1, a covariance function C describes the covariance between sample i and sample j in terms of the D features

$$\boldsymbol{K}_{ij} = C\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right), \qquad \text{(D.12)}$$

where $\boldsymbol{x}_i = \boldsymbol{X}_{i,:}$ and $\boldsymbol{x}_j = \boldsymbol{X}_{j,:}$.

$$\boldsymbol{K}_{\boldsymbol{\theta} ij} = e^{a} \exp\left(-\frac{\sum_d \left(\boldsymbol{x}_{id} - \boldsymbol{x}_{jd}\right)^2}{2e^{b}}\right), \qquad \text{(D.13)}$$

where $\boldsymbol{\theta} = \{a, b\}$ and $\sigma^2 = e^a$ and $l = e^{\frac{b}{2}}$ are known as amplitude and scale. The derivatives are

$$\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta} ij}}{\partial a} = e^a \exp\left(-\frac{\sum_d (\boldsymbol{x}_{id} - \boldsymbol{x}_{jd})^2}{2e^b}\right) \tag{D.14}$$

$$\frac{\partial \boldsymbol{K}_{\boldsymbol{\theta} ij}}{\partial b} = e^a \exp\left(-\frac{\sum_d (\boldsymbol{x}_{id} - \boldsymbol{x}_{jd})^2}{2e^b}\right) \frac{\sum_d (\boldsymbol{x}_{id} - \boldsymbol{x}_{jd})^2}{2e^b} \tag{D.15}$$

D.3 Standard errors

Indicating with $\boldsymbol{I}$ the Fisher information matrix for a gaussian likelihood function for variance parameters we have

$$\boldsymbol{I}_{mn} = \frac{1}{2}\text{tr}\left(\boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_m} \boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \frac{\partial \boldsymbol{K}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}_n}\right) \tag{D.16}$$

The covariance matrix between parameters $\boldsymbol{\Sigma}$ is the inverse of the Fisher information, $\boldsymbol{\Sigma} = \boldsymbol{I}^{-1}$ and the standard errors its diagonal, $\text{ste}\left(\hat{\boldsymbol{\theta}}_m\right) = \boldsymbol{\Sigma}_{mm}$.

D.4 Supplementary information for the analysis in BluePrint WP10

D.4.1 Molecular assays and data preprocessing

Blood collection and cell isolation were perfumed at the University of Cambridge (UoC). The produced monocyte population had purity 95%. As described in more detail below, the subsequent molecular assays, quality control (QC) steps and data processing were performed in different research institutes, including the Wellcome Trust Sanger Institute (WTSI), the Max Planck Institute for Molecular Genetics (MPIMG), the University College London (UCL), the Nijmegen Centre for Molecular Life Sciences (NCMLS) and the European Bioinformatics Institute (EBI). I have not taken part to the collection and the preprocessing of the data. The analyses I performed are described in Section 5.3.2, for which I considered the set of 158 individuals for which all five molecular layers (genotype, DNA methylation, H3K4me1, H3K27ac and expression) had passed all QC steps.

WGS data. DNA extraction was performed at the UoC. Sequencing (100bp pair-ended; HiSeq 2000/2500, Illumina; 7.05x coverage), alignment to reference genome

(using BWA, Li and Durbin, 2009), variant calling (SAMtools / bcftools, Li, 2011) and QC steps were performed at the WTSI. In the analyses discussed in the next sections we considered a set of 5,237,919 common variants (MAF $> 4\%$).

RNA-sequencing. Library preparation and sequencing (100 bp single end (SE); V3 chemistry, HiSeq 2000, Illumina) were performed at MPIMG. Pre-alignment QC steps, alignment (with STAR, Dobin et al., 2013) and gene-level quantification (DESeq2, Love et al., 2014) was performed at WTSI. Only the 16,577 genes with at least 10 read counts in at least 50% of the samples were considered for further analysis.

Methylation data. Data generation with 450K array and data processing were carried out at the UCL. CpG probes (i) with median P value ≥ 0.01 in at least one sample, (ii) with bead count ≤ 3 in at least three samples, (iii) mapping to sex chromosomes or to multiple locations, (iv) with common SNP (MAF ≥ 0.05) within 2 bp from the CpG site were excluded. This led to a set of 440,905 CpG sites that we considered for further analysis. Methylation of CpG sites was expressed using M-values, defined as the $\log_2$ ratio of the intensities of methylated and unmethylated probes.

Histone modification. Chromatin immunoprecipitation (ChIP) assay for H3K27ac/H3K4me1 histone marks and sequencing were performed in different sequencing centres using different protocols. Monocytes 1-49/1-48 are were processed at teh NCMLS (Illumina HiSeq 2000 at 43bp SE reads) while monocytes 50-162/49-172 at the WTSI (Illumina HiSeq 2000 at 50bp SE reads). Alignment to reference genome (BWA, Li and Durbin, 2009), QC steps, peak calling (using MACS2, FDR $\leq 1\%$) and peak intensity quantification were performed at the EBI. For each histone modification marker the reference peak set was obtained by (i) considering the union of significant peaks across all individuals and cell-types and (ii) merging overlapping peaks. This procedure led to the identification of 64,843 and 39,815 large peaks ($\geq$ 100kb) for H3K27ac and H3K4me1 histone modifications, respectively. Peak intensities were expressed as $\log_2$ of the total number of read counts within the peak per million base pairs and were normalised with respect to the total number of reads in the library.

Data were corrected for batch effects and technical confounding using ComBat (Leek et al., 2012).

D.4.2 Accounting for sample heterogeneity

To assess whether the correction using PEER was sufficient to eliminate this potential confounding, we proceeded as described in the following. Let N denote the number of individuals, $\boldsymbol{y} \in \mathbb{R}^N$ the normalised gene-expression vector, $\boldsymbol{K}_g \in \mathbb{R}^{N \times N}$ the global RRM and $\boldsymbol{K}_{\text{cis}} \in \mathbb{R}^{N \times N}$ a local RRM based on the genetic variants in the *cis* region (1Mb either side from the gene body). For each gene, we considered the model

$$\boldsymbol{y} \sim \mathcal{N}\left(\underbrace{\mathbf{1}\mu}_{\text{intercept term}} \quad , \quad \underbrace{\sigma^2_{\text{cis}} \boldsymbol{K}_{\text{cis}}}_{\text{cis component}} \quad + \quad \underbrace{\sigma^2_g \boldsymbol{K}_g}_{\text{relatedness component}} \quad + \underbrace{\sigma^2_e \boldsymbol{I}}_{\text{noise}} \right), \tag{D.17}$$

where $\mathbf{1}\mu \in \mathbb{R}^N$ is an intercept term. As discussed in Section 2.3.4, this model can be used to estimate the proportion of phenotypic variance explained jointly by all *cis* variants. We used the same approach to estimate the variance explained by *cis* CpG sites, *cis* H3K27ac peaks and *cis* H3K4me1 peaks by using a local relatedness matrices based on the corresponding *cis* epigenetic features. Epigenetic features were quantile-normalised to a unit variance normal distribution before computing the local (epigenetic) relatedness matrices. Denoting with $\boldsymbol{Z}$ denotes the $N \times G$ gene expression matrix for N individuals and all G genes, let us consider the expression heterogeneity (EH) covariance as $\boldsymbol{K}_h = \boldsymbol{Z}\boldsymbol{Z}^T \in \mathbb{R}^{N \times N}$. Under the assumption that genome-wide expression heterogeneity can be used as a tag for technical confounding, a common strategy in eQTL mapping (Kang et al., 2008a; Fusi et al., 2012), we introduced $\boldsymbol{K}_h$ in the model and re-estimated the cis variance components for the genetic and the three epigenetic layers. We found that the model accounting for EH yielded substantially lower epigenome variance estimates, whereas the cis genetic variances were robust (**Fig. D.1**). We thereby decided to account for EH in all subsequent analyses.

D.4.3 Supplementary Figures

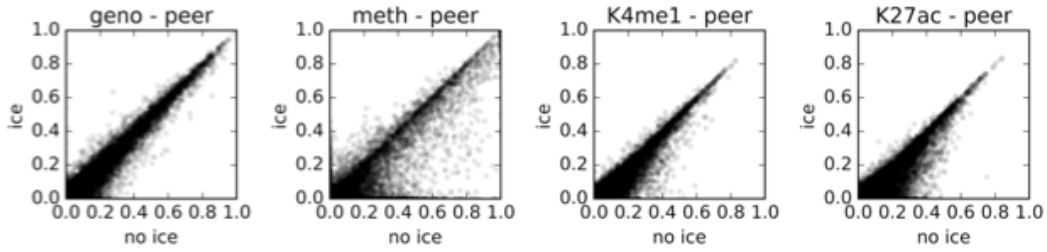


Figure D.1: **Comparison of variance component estimates either accounting or not for expression heterogeneity.** Compared are the estimated proportion of variance explained by cis genetics, cis methylation and cis histone marks when either considering a model that accounts for expression heterogeneity (y-axis) or standard linear mixed model (x-axis). While genetic estimates were robust, epigenetic variance estimates from the model that accounts for heterogeneity were substantially lower.

E | Publications

1. **Casale FP**, Horta D, Rakitsch B, Stegle O. Joint genetic analysis using variant sets reveals polygenic gene-context interactions. PLoS genetics. 2017 Apr 20;13(4):e1006693.

2. Schor IE, Degner JF, Harnett D, Cannavo E, **Casale FP**, Shim H, Garfield D, Birney E, Stephens M, Stegle O, Furlong E. Common genetic variants determine promoter shape with implications for robustness and expression noise. Nature Genetics (2017).

3. Baud A, Mulligan MK, **Casale FP**, Ingels JF, Bohl CJ, Callebert J, Launay JM, Krohn J, Legarra A, Williams RW, Stegle O. Genetic Variation in the Social Environment Contributes to Health and Disease. PLoS genetics. 2017 Jan 25;13(1):e1006498.

4. Ecker S, Chen L, Pancaldi V, Bagger FO, Fernandez JM, de Santa Pau EC, Juan D, Mann A, Watt S, **Casale FP**, Sidiropoulos N et al. Genome-wide Analysis of Differential Transcriptional and Epigenetic Variability Across Human Immune Cell Types. Genome Biology. 2017 Jan 26.

5. Cannavo E, Koelling N, Harnett D, Garfield D, **Casale FP**, Ciglar L, Gustafson HE, Viales RR, Marco-Ferreres R, Degner JF et al. Genetic variants regulating expression levels and isoform diversity during embryogenesis. Nature. 2016 Dec 26.

6. Chen L*, Ge B*, **Casale FP***, Vasquez L*, Kwan T, Garrido-Martin D, Watt S, Yan Y, Kundu K, Ecker S, Datta A et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016 Nov 17;167(5):1398-414.

7. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, Konkel MK et al. An integrated map of

structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75-81.

8. **Casale FP***, Rakitsch B*, Lippert C, Stegle O. Efficient set tests for the genetic analysis of correlated traits. Nature methods. 2015 Aug 1;12(8):755-8.

9. Dubin MJ, Zhang P, Meng D, Remigereau MS, Osborne EJ, **Casale FP**, Drewe P, Kahles A, Jean G, Vilhjalmsson B, Jagoda J et al. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife. 2015 May 5;4:e05255.

10. Buettner F, Natarajan KN, **Casale FP**, Proserpio V, Scialdone A, Theis FJ, Teichmann SA, Marioni JC, Stegle O. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nature biotechnology. 2015 Feb 1;33(2):155-60.

In review / submitted

12. Menden MP*, **Casale FP***, Stephan J, Bignell GR, Iorio F, McDermott U, Garnett MJ, Saez-Rodriguez J, Stegle O. The germline genetic component of drug sensitivity in cancer cell lines, submitted

13. Kilpinen H, Goncalves A et al. Common genetic variation drives molecular heterogeneity in human iPSCs, bioRxiv doi:10.1101/055160, in review

14. Lippert C*, **Casale FP***, Rakitsch B, Stegle O. LIMIX: genetic analysis of multiple traits. BioRxiv. 2014 Jan 1:003905.

* joint first author.

Bibliography

1000 Genomes Project Consortium et al. (2015). "A global reference for human genetic variation". *Nature* 526.7571, pp. 68–74.

1000 Genomes Project Consortium et al. (2010). "A map of human genome variation from population-scale sequencing". *Nature* 467.7319, pp. 1061–1073.

1000 Genomes Project Consortium et al. (2012). "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491.7422, pp. 56–65.

Alberts, Bruce, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, and Peter Walter (2014). *Molecular Biology of the Cell*. 6th ed. Garland Science.

Altmüller, Janine, Lyle J Palmer, Guido Fischer, Hagen Scherb, and Matthias Wjst (2001). "Genomewide scans of complex human diseases: true linkage is hard to find". *The American Journal of Human Genetics* 69.5, pp. 936–950.

Andreassi, Maria Grazia (2009). "Metabolic syndrome, diabetes and atherosclerosis: influence of gene-environment interaction." *Mutat. Res.* 667.1-2, pp. 35–43.

Aschard, Hugues, Bjarni J Vilhjálmsson, Nicolas Greliche, Pierre-Emmanuel Morange, David-Alexandre Trégouët, and Peter Kraft (2014). "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies." *Am. J. Hum. Genet.* 94.5, pp. 662–76.

Astle, William and David J Balding (2009). "Population structure and cryptic relatedness in genetic association studies". *Statistical Science*, pp. 451–471.

Atwell, Susanna, Yu S Huang, Bjarni J Vilhjálmsson, Glenda Willems, Matthew Horton, Yan Li, Dazhe Meng, Alexander Platt, Aaron M Tarone, et al. (2010). "Genome-

wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines." *Nature* 465.7298, pp. 627–31.

Battle, Alexis, Sara Mostafavi, Xiaowei Zhu, James B Potash, Myrna M Weissman, Courtney McCormick, Christian D Haudenschild, Kenneth B Beckman, Jianxin Shi, et al. (2014). "Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals." *Genome Res.* 24.1, pp. 14–24.

Bauchet, Marc, Brian McEvoy, Laurel N Pearson, Ellen E Quillen, Tamara Sarkisian, Kristine Hovhannesyan, Ranjan Deka, Daniel G Bradley, and Mark D Shriver (2007). "Measuring European population stratification with microarray genotype data." *Am. J. Hum. Genet.* 80.5, pp. 948–56.

Baud, Amelie, Megan K Mulligan, Francesco Paolo Casale, Jesse F Ingels, Casey J Bohl, Jacques Callebert, Jean-Marie Launay, Jon Krohn, Andres Legarra, Robert W Williams, et al. (2017). "Genetic variation in the social environment contributes to health and disease". *PLoS genetics* 13.1, e1006498.

Baud, Amelie, Victor Guryev, Oliver Hummel, Martina Johannesson, and Jonathan Flint (2014). "Genomes and phenomes of a population of outbred rats and its progenitors." *Sci Data* 1, p. 140011.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300.

Benner, Christian, Chris C A Spencer, Aki S Havulinna, Veikko Salomaa, Samuli Ripatti, and Matti Pirinen (2016). "FINEMAP: efficient variable selection using summary data from genome-wide association studies." *Bioinformatics* 32.10, pp. 1493–501.

Bijma, P (2014). "The quantitative genetics of indirect genetic effects: a selective review of modelling issues." *Heredity (Edinb)* 112.1, pp. 61–9.

Birney, Ewan and Nicole Soranzo (2015). "Human genomics: The end of the start for population sequencing." *Nature* 526.7571, pp. 52–3.

Bloom, Joshua S, Ian M Ehrenreich, Wesley T Loo, Thúy-Lan Võ Lite, and Leonid Kruglyak (2013). "Finding the sources of missing heritability in a yeast cross". *Nature* 494.7436, pp. 234–237.

Bolormaa, Sunduimijid, Jennie E Pryce, Antonio Reverter, Yuandan Zhang, William Barendse, Kathryn Kemper, Bruce Tier, Keith Savin, Ben J Hayes, and Michael E Goddard (2014). "A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle." *PLoS Genet.* 10.3, e1004198.

Bottolo, Leonardo, Marc Chadeau-Hyam, David I Hastie, Tanja Zeller, Benoit Liquet, Paul Newcombe, Loic Yengo, Philipp S Wild, Arne Schillert, et al. (2013). "GUESSing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm." *PLoS Genet.* 9.8, e1003657.

Brent, Richard P. (1971). "An algorithm with guaranteed convergence for finding a zero of a function". *The Computer Journal* 14.4, pp. 422–425.

Broadaway, K Alaine, Richard Duncan, Karen N Conneely, Lynn M Almli, Bekh Bradley, Kerry J Ressler, and Michael P Epstein (2015). "Kernel Approach for Modeling Interaction Effects in Genetic Association Studies of Complex Quantitative Traits." *Genet. Epidemiol.* 39.5, pp. 366–75.

Brosius, Jürgen (2009). "The fragmented gene." *Ann. N. Y. Acad. Sci.* 1178, pp. 186–93.

Brown, Brielin C, Alkes L Price, Nikolaos A Patsopoulos, and Noah Zaitlen (2016). "Local Joint Testing Improves Power and Identifies Hidden Heritability in Association Studies." *Genetics* 203.3, pp. 1105–16.

Bůžková, Petra, Thomas Lumley, and Kenneth Rice (2011). "Permutation and parametric bootstrap tests for gene–gene and gene–environment interactions". *Annals of human genetics* 75.1, pp. 36–45.

Bulik-Sullivan, Brendan K, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale (2015). "LD Score regression distinguishes confounding from polygenicity in genome-wide association studies." *Nat. Genet.* 47.3, pp. 291–5.

Burdett, T, PN Hall, E Hastings, LA Hindorff, HA Junkins, et al. (2015). "The NHGRI-EBI Catalog of published genome-wide association studies". *Available at: www. ebiacuk/gwas*.

Burton, Paul R, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls". *Nature* 447.7145, pp. 661–678.

Bush, William S and Jason H Moore (2012). "Chapter 11: Genome-wide association studies." *PLoS Comput. Biol.* 8.12, e1002822.

Cai, Na, Simon Chang, Yihan Li, Qibin Li, Jingchu Hu, Jieqin Liang, Li Song, Warren Kretzschmar, Xiangchao Gan, et al. (2015). "Molecular signatures of major depression." *Curr. Biol.* 25.9, pp. 1146–56.

Cannavò, Enrico, Nils Koelling, Dermot Harnett, David Garfield, Francesco P Casale, Lucia Ciglar, Hilary E Gustafson, Rebecca R Viales, Raquel Marco-Ferreres, Jacob F Degner, et al. (2016). "Genetic variants regulating expression levels and isoform diversity during embryogenesis". *Nature*.

Carninci, Piero, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, Martin S Taylor, Pär G Engström, et al. (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." *Nat. Genet.* 38.6, pp. 626–35.

Casale, Francesco Paolo, Barbara Rakitsch, Christoph Lippert, and Oliver Stegle (2015). "Efficient set tests for the genetic analysis of correlated traits." *Nat. Methods* 12.8, pp. 755–8.

Chatterjee, Nilanjan, Zeynep Kalaylioglu, Roxana Moslehi, Ulrike Peters, and Sholom Wacholder (2006). "Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions." *Am. J. Hum. Genet.* 79.6, pp. 1002–16.

Chen, Han, James B Meigs, and Josée Dupuis (2013). "Sequence kernel association test for quantitative traits in family samples." *Genet. Epidemiol.* 37.2, pp. 196–204.

Chen, Lu, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, Ying Yan, Kousik Kundu, Simone Ecker, et al. (2016). "Genetic drivers of epigenetic and transcriptional variation in human immune cells". *Cell* 167.5, pp. 1398–1414.

Chen, Wenan, Beth R Larrabee, Inna G Ovsyannikova, Richard B Kennedy, Iana H Haralambieva, Gregory A Poland, and Daniel J Schaid (2015). "Fine mapping causal variants with an approximate Bayesian method using marginal test statistics". *Genetics* 200.3, pp. 719–736.

Chiba-Falek, Ornit, Colton Linnertz, John Guyton, Stephen D Gardner, Allen D Roses, Jeanette J McCarthy, and Keyur Patel (2012). "Pleiotropy and allelic heterogeneity in the TOMM40-APOE genomic region related to clinical and metabolic features of hepatitis C infection." *Hum. Genet.* 131.12, pp. 1911–20.

Clark, Samuel A and Julius van der Werf (2013). "Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values." *Methods Mol. Biol.* 1019, pp. 321–30.

Cohen, Jonathan C, Robert S Kiss, Alexander Pertsemlidis, Yves L Marcel, Ruth McPherson, and Helen H Hobbs (2004). "Multiple rare alleles contribute to low plasma levels of HDL cholesterol." *Science* 305.5685, pp. 869–72.

Conneely, Karen N and Michael Boehnke (2007). "So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests." *Am. J. Hum. Genet.* 81.6, pp. 1158–68.

Corradin, Olivia, Alina Saiakhova, Batool Akhtar-Zaidi, Lois Myeroff, Joseph Willis, Richard Cowper-Sal lari, Mathieu Lupien, Sanford Markowitz, and Peter C Scacheri (2014). "Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits." *Genome Res.* 24.1, pp. 1–13.

Creyghton, Menno P, Albert W Cheng, G Grant Welstead, Tristan Kooistra, Bryce W Carey, Eveline J Steine, Jacob Hanna, Michael A Lodato, Garrett M Frampton, et al. (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state." *Proc. Natl. Acad. Sci. U.S.A.* 107.50, pp. 21931–6.

Dahl, Andrew, Valentina Iotchkova, Amelie Baud, Åsa Johansson, Ulf Gyllensten, Nicole Soranzo, Richard Mott, Andreas Kranis, and Jonathan Marchini (2016). "A multiple-phenotype imputation method for genetic studies." *Nat. Genet.* 48.4, pp. 466–72.

Dahl, Andy, Victoria Hore, Valentina Iotchkova, and Jonathan Marchini (2013). "Network inference in matrix-variate Gaussian models with non-independent noise". *arXiv preprint arXiv:1312.1622*.

Davies, Robert B (1980). "Algorithm AS 155: The distribution of a linear combination of χ 2 random variables". *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29.3, pp. 323–333.

Davis, Joe R, Laure Fresard, David A Knowles, Mauro Pala, Carlos D Bustamante, Alexis Battle, and Stephen B Montgomery (2016). "An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants." *Am. J. Hum. Genet.* 98.1, pp. 216–24.

Dehghan, Abbas, Josée Dupuis, Maja Barbalic, Joshua C Bis, Gudny Eiriksdottir, Chen Lu, Niina Pellikka, Henri Wallaschofski, Johannes Kettunen, et al. (2011). "Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels." *Circulation* 123.7, pp. 731–8.

Delaneau, Olivier, Jonathan Marchini, and and (2014). "Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel." *Nat Commun* 5, p. 3934.

Dempster, Arthur P, Donald B Rubin, and Robert K Tsutakawa (1981). "Estimation in covariance components models". *Journal of the American Statistical Association* 76.374, pp. 341–353.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Devlin, B and N Risch (1995). "A comparison of linkage disequilibrium measures for fine-scale mapping." *Genomics* 29.2, pp. 311–22.

Devlin, B and K Roeder (1999). "Genomic control for association studies." *Biometrics* 55.4, pp. 997–1004.

Dick, Danielle M (2011). "Gene-environment interaction in psychological traits and disorders." *Annu Rev Clin Psychol* 7, pp. 383–409.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." *Bioinformatics* 29.1, pp. 15–21.

Dominicus, Annica, Anders Skrondal, Håkon K Gjessing, Nancy L Pedersen, and Juni Palmgren (2006). "Likelihood ratio tests in behavioral genetics: problems and solutions". *Behavior genetics* 36.2, pp. 331–340.

Dubin, Manu J, Pei Zhang, Dazhe Meng, Marie-Stanislas Remigereau, Edward J Osborne, Francesco Paolo Casale, Philipp Drewe, André Kahles, Geraldine Jean, et al. (2015a). "DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation." *Elife* 4, e05255.

Dubin, Manu J, Pei Zhang, Dazhe Meng, Marie-Stanislas Remigereau, Edward J Osborne, Francesco Paolo Casale, Philipp Drewe, André Kahles, Geraldine Jean, Bjarni Vilhjálmsson, et al. (2015b). "DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation". *Elife* 4, e05255.

Dupuis, Josée, Claudia Langenberg, Inga Prokopenko, Richa Saxena, Nicole Soranzo, Anne U Jackson, Eleanor Wheeler, Nicole L Glazer, Nabila Bouatia-Naji, et al. (2010). "New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk." *Nat. Genet.* 42.2, pp. 105–16.

Ehret, Georg B, David Lamparter, Clive J Hoggart, John C Whittaker, Jacques S Beckmann, and Zoltán Kutalik (2012). "A multi-SNP locus-association method reveals a substantial fraction of the missing heritability." *Am. J. Hum. Genet.* 91.5, pp. 863–71.

Eichler, Evan E, Jonathan Flint, Greg Gibson, Augustine Kong, Suzanne M Leal, Jason H Moore, and Joseph H Nadeau (2010). "Missing heritability and strategies for finding the underlying causes of complex disease." *Nat. Rev. Genet.* 11.6, pp. 446–50.

ENCODE Project Consortium et al. (2004). "The ENCODE (ENCyclopedia of DNA elements) project". *Science* 306.5696, pp. 636–640.

Ernst, Jason and Manolis Kellis (2010). "Discovery and characterization of chromatin states for systematic annotation of the human genome." *Nat. Biotechnol.* 28.8, pp. 817–25.

Eu-Ahsunthornwattana, Jakris, E Nancy Miller, Michaela Fakiola, Selma M B Jeronimo, Jenefer M Blackwell, and Heather J Cordell (2014). "Comparison of methods to account for relatedness in genome-wide association studies with family-based data." *PLoS Genet.* 10.7, e1004445.

Ewens, W J and R S Spielman (1995). "The transmission/disequilibrium test: history, subdivision, and admixture." *Am. J. Hum. Genet.* 57.2, pp. 455–64.

Fairfax, Benjamin P, Peter Humburg, Seiko Makino, Vivek Naranbhai, Daniel Wong, Evelyn Lau, Luke Jostins, Katharine Plant, Robert Andrews, et al. (2014). "Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression." *Science* 343.6175, p. 1246949.

Falconer, Douglas S and Trudy FC Mackay (1996). "Introduction to quantitative genetics". *Harlow, UK: Longman.*

Ferreira, Manuel A R and Shaun M Purcell (2009). "A multivariate test of association." *Bioinformatics* 25.1, pp. 132–3.

Fisher, RA (1918). "The Correlation Between Relatives on the Supposition of Mendelian Inheritance". *Transactions of the Royal Society of Edinburgh* 52.02, pp. 399–433.

Fisher, Ronald A (1919). "XV.—The correlation between relatives on the supposition of Mendelian inheritance." *Transactions of the royal society of Edinburgh* 52.02, pp. 399–433.

Fisher, Ronald A and Winifred A Mackenzie (1923). "Studies in crop variation. II. The manurial response of different potato varieties". *The Journal of Agricultural Science* 13.03, pp. 311–320.

Fisher, Ronald Aylmer (1921). "Studies in Crop Variation. I. An examination of the yield of dressed grain from Broadbalk". *The Journal of Agricultural Science* 11.02, pp. 107–135.

Flutre, Timothée, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens (2013). "A statistical framework for joint eQTL analysis in multiple tissues." *PLoS Genet.* 9.5, e1003486.

Fortune, Mary D, Hui Guo, Oliver Burren, Ellen Schofield, Neil M Walker, Maria Ban, Stephen J Sawcer, John Bowes, Jane Worthington, et al. (2015). "Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls." *Nat. Genet.* 47.7, pp. 839–46.

Foulley, Jean-Louis and David A Van Dyk (2000). "The PX-EM algorithm for fast stable fitting of Henderson's mixed model". *Genetics Selection Evolution* 32.2, p. 1.

Francesconi, Mirko and Ben Lehner (2014). "The effects of genetic variation on gene expression dynamics during development." *Nature* 505.7482, pp. 208–11.

Frazer, Kelly A, Sarah S Murray, Nicholas J Schork, and Eric J Topol (2009). "Human genetic variation and its contribution to complex traits." *Nat. Rev. Genet.* 10.4, pp. 241–51.

Furlotte, Nicholas A and Eleazar Eskin (2015). "Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model." *Genetics* 200.1, pp. 59–68.

Fusi, Nicoló, Oliver Stegle, and Neil D Lawrence (2012). "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies." *PLoS Comput. Biol.* 8.1, e1002330.

Fusi, Nicolo, Christoph Lippert, Neil D Lawrence, and Oliver Stegle (2014a). "Warped linear mixed models for the genetic analysis of transformed phenotypes." *Nat Commun* 5, p. 4890.

Fusi, Nicolo, Christoph Lippert, Neil D Lawrence, and Oliver Stegle (2014b). "Warped linear mixed models for the genetic analysis of transformed phenotypes". *Nature communications* 5.

Gagneur, Julien, Oliver Stegle, Chenchen Zhu, Petra Jakob, Manu M Tekkedil, Raeka S Aiyar, Ann-Kathrin Schuon, Dana Pe'er, and Lars M Steinmetz (2013). "Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype." *PLoS Genet.* 9.9, e1003803.

Galton, Francis (1869). *Hereditary genius: An inquiry into its laws and consequences.* Vol. 27. Macmillan.

Gao, Xiaoyi, Joshua Starmer, and Eden R Martin (2008). "A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms." *Genet. Epidemiol.* 32.4, pp. 361–9.

Gauderman, W James, Pingye Zhang, John L Morrison, and Juan Pablo Lewinger (2013). "Finding novel genes by testing G×E interactions in a genome-wide association study." *Genet. Epidemiol.* 37.6, pp. 603–13.

Gauderman, W James, Cassandra Murcray, Frank Gilliland, and David V Conti (2007). "Testing association between disease and multiple SNPs in a candidate gene". *Genet Epidemiol* 31.5, pp. 383–95.

Gaunt, Tom R, Hashem A Shihab, Gibran Hemani, Josine L Min, Geoff Woodward, Oliver Lyttleton, Jie Zheng, Aparna Duggirala, Wendy L McArdle, et al. (2016). "Systematic identification of genetic influences on methylation across the human life course." *Genome Biol.* 17, p. 61.

Giambartolomei, Claudia, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol (2014). "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics." *PLoS Genet.* 10.5, e1004383.

Gibbs, Richard A, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. (2003). "The international HapMap project". *Nature* 426.6968, pp. 789–796.

Gilmour, Arthur R, BJ Gogel, BR Cullis, R Thompson, D Butler, et al. (2009). "ASReml user guide release 3.0". *VSN International Ltd, Hemel Hempstead, UK.*

Gilmour, Arthur R, Robin Thompson, and Brian R Cullis (1995). "Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models". *Biometrics*, pp. 1440–1450.

Goncalo R Abecasis, and, Adam Auton, Lisa D Brooks, Mark A DePristo, Richard M Durbin, Robert E Handsaker, Hyun Min Kang, Gabor T Marth, and Gil A McVean (2012). "An integrated map of genetic variation from 1,092 human genomes." *Nature* 491.7422, pp. 56–65.

Gormley, Padhraig, Verneri Anttila, Bendik S Winsvold, Priit Palta, Tonu Esko, Tune H Pers, Kai-How Farh, Ester Cuenca-Leon, Mikko Muona, et al. (2016). "Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine." *Nat. Genet.* 48.8, pp. 856–66.

GPy (since 2012). *GPy: A Gaussian process framework in python.* `http://github.com/SheffieldML/GPy`.

Greven, Sonja, Ciprian M Crainiceanu, Helmut Küchenhoff, and Annette Peters (2012). "Restricted likelihood ratio testing for zero variance components in linear mixed models". *Journal of Computational and Graphical Statistics*.

Griffiths, Anthony JF (2005). *An introduction to genetic analysis*. Macmillan.

Groeneveld, E (1994). "A reparameterization to improve numerical optimization in multivariate REML (co) variance component estimation". *Genetics Selection Evolution* 26.6, p. 1.

Grubert, Fabian, Judith B Zaugg, Maya Kasowski, Oana Ursu, Damek V Spacek, Alicia R Martin, Peyton Greenside, Rohith Srivas, Doug H Phanstiel, et al. (2015). "Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions." *Cell* 162.5, pp. 1051–65.

GTEx Consortium et al. (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". *Science* 348.6235, pp. 648–660.

Gusella, J F (1984). "Genetic linkage of the Huntington's disease gene to a DNA marker." *Can J Neurol Sci* 11.4, pp. 421–5.

Gusev, Alexander, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, et al. (2014). "Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases." *Am. J. Hum. Genet.* 95.5, pp. 535–52.

Gusev, Alexander, Gaurav Bhatia, Noah Zaitlen, Bjarni J Vilhjalmsson, Dorothée Diogo, Eli A Stahl, Peter K Gregersen, Jane Worthington, Lars Klareskog, et al. (2013). "Quantifying missing heritability at known GWAS loci." *PLoS Genet.* 9.12, e1003993.

Harville, David A (1974). "Bayesian inference for variance components using only error contrasts". *Biometrika* 61.2, pp. 383–385.

Harville, David A (1998). "Matrix algebra from a statistician's perspective". *Technometrics* 40.2, pp. 164–164.

Hayes, B J, P M Visscher, and M E Goddard (2009). "Increased accuracy of artificial selection by using the realized relationship matrix." *Genet Res (Camb)* 91.1, pp. 47–60.

Henderson, Charles R (1950). "Estimation of genetic parameters (Abstract)". *The Annals of Mathematical Statistics*, pp. 309–310.

Henderson, Charles R, Oscar Kempthorne, Shayle R Searle, and CM Von Krosigk (1959). "The estimation of environmental and genetic trends from records subject to culling". *Biometrics* 15.2, pp. 192–218.

Henderson, CR (1984). "1984-Guelph".

Hoerl, Arthur E and Robert W Kennard (1970). "Ridge regression: Biased estimation for nonorthogonal problems". *Technometrics* 12.1, pp. 55–67.

Hoffman, Gabriel E (2013). "Correcting for population structure and kinship using the linear mixed model: theory and extensions." *PLoS ONE* 8.10, e75707.

Hon, Gary C, R David Hawkins, and Bing Ren (2009). "Predictive chromatin signatures in the mammalian genome." *Hum. Mol. Genet.* 18.R2, R195–201.

Horton, Matthew W, Glenda Willems, Eriko Sasaki, Maarten Koornneef, and Magnus Nordborg (2016). "The genetic architecture of freezing tolerance varies across the range of Arabidopsis thaliana." *Plant Cell Environ.*

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R Abecasis (2012). "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing." *Nat. Genet.* 44.8, pp. 955–9.

Howie, Bryan N, Peter Donnelly, and Jonathan Marchini (2009). "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." *PLoS Genet.* 5.6, e1000529.

Huang, Jie, Bryan Howie, Shane McCarthy, Yasin Memari, Klaudia Walter, Josine L Min, Petr Danecek, Giovanni Malerba, Elisabetta Trabetti, et al. (2015). "Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel." *Nat Commun* 6, p. 8111.

International HapMap Consortium et al. (2005). "A haplotype map of the human genome". *Nature* 437.7063, pp. 1299–1320.

Jakobsson, Mattias, Sonja W Scholz, Paul Scheet, J Raphael Gibbs, Jenna M VanLiere, Hon-Chung Fung, Zachary A Szpiech, James H Degnan, Kai Wang, et al. (2008). "Genotype, haplotype and copy-number variation in worldwide human populations." *Nature* 451.7181, pp. 998–1003.

Jiang, C and Z B Zeng (1995). "Multiple trait analysis of genetic mapping for quantitative trait loci." *Genetics* 140.3, pp. 1111–27.

Jiao, Shuo, Li Hsu, Stéphane Bézieau, Hermann Brenner, Andrew T Chan, Jenny Chang-Claude, Loic Le Marchand, Mathieu Lemire, Polly A Newcomb, et al. (2013). "SBERIA: set-based gene-environment interaction test for rare and common variants in complex diseases." *Genet. Epidemiol.* 37.5, pp. 452–64.

Jones, Eric, Travis Oliphant, Pearu Peterson, et al. (2001). *SciPy: Open source scientific tools for Python.*

Kang, Eun Yong, Buhm Han, Nicholas Furlotte, Jong Wha J Joo, Diana Shih, Richard C Davis, Aldons J Lusis, and Eleazar Eskin (2014). "Meta-analysis identifies

gene-by-environment interactions as demonstrated in a study of 4,965 mice." *PLoS Genet.* 10.1, e1004022.

Kang, Hyun Min, Chun Ye, and Eleazar Eskin (2008a). "Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots." *Genetics* 180.4, pp. 1909–25.

Kang, Hyun Min, Noah A Zaitlen, Claire M Wade, Andrew Kirby, David Heckerman, Mark J Daly, and Eleazar Eskin (2008b). "Efficient control of population structure in model organism association mapping." *Genetics* 178.3, pp. 1709–23.

Kang, Hyun Min, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-Yee Kong, Nelson B Freimer, Chiara Sabatti, and Eleazar Eskin (2010). "Variance component model to account for sample structure in genome-wide association studies." *Nat. Genet.* 42.4, pp. 348–54.

Kawakatsu, Taiji, Shao-Shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, Joseph R Nery, Cesar Barragan, et al. (2016). "Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions." *Cell* 166.2, pp. 492–505.

Khor, Bernard, Agnès Gardet, and Ramnik J Xavier (2011). "Genetics and pathogenesis of inflammatory bowel disease." *Nature* 474.7351, pp. 307–17.

Kim, Yun Kyoung, Youngdoe Kim, Mi Yeong Hwang, Kazuro Shimokawa, Sungho Won, Norihiro Kato, Yasuharu Tabara, Mitsuhiro Yokota, Bok-Ghee Han, et al. (2014). "Identification of a genetic variant at 2q12.1 associated with blood pressure in East Asians by genome-wide scan including gene-environment interactions." *BMC Med. Genet.* 15, p. 65.

Klaudia Walter, and, Josine L Min, Jie Huang, Lucy Crooks, Yasin Memari, Shane McCarthy, John R B Perry, ChangJiang Xu, Marta Futema, et al. (2015). "The UK10K project identifies rare variants in health and disease." *Nature* 526.7571, pp. 82–90.

Klein, Robert J, Caroline Zeiss, Emily Y Chew, Jen-Yue Tsai, Richard S Sackler, Chad Haynes, Alice K Henning, John Paul SanGiovanni, Shrikant M Mane, et al. (2005). "Complement factor H polymorphism in age-related macular degeneration." *Science* 308.5720, pp. 385–9.

Kohli, Martin A, Susanne Lucae, Philipp G Saemann, Mathias V Schmidt, Ayse Demirkan, Karin Hek, Darina Czamara, Michael Alexander, Daria Salyakina, et al. (2011). "The neuronal transporter gene SLC6A15 confers risk to major depression." *Neuron* 70.2, pp. 252–65.

Koishi, Ryuta, Yosuke Ando, Mitsuru Ono, Mitsuru Shimamura, Hiroaki Yasumo, Toshihiko Fujiwara, Hiroyoshi Horikoshi, and Hidehiko Furukawa (2002). "Angptl3 regulates lipid metabolism in mice." *Nat. Genet.* 30.2, pp. 151–7.

Korte, Arthur, Bjarni J Vilhjálmsson, Vincent Segura, Alexander Platt, Quan Long, and Magnus Nordborg (2012). "A mixed-model approach for genome-wide association studies of correlated traits in structured populations." *Nat. Genet.* 44.9, pp. 1066–71.

Kostem, Emrah and Eleazar Eskin (2013). "Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions." *Am. J. Hum. Genet.* 92.4, pp. 558–64.

Kraja, Aldi T, Dhananjay Vaidya, James S Pankow, Mark O Goodarzi, Themistocles L Assimes, Iftikhar J Kullo, Ulla Sovio, Rasika A Mathias, Yan V Sun, et al. (2011). "A bivariate genome-wide approach to metabolic syndrome: STAMPEED consortium." *Diabetes* 60.4, pp. 1329–39.

Kundaje, Anshul, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. (2015). "Integrative analysis of 111 reference human epigenomes". *Nature* 518.7539, pp. 317–330.

Laird, Nan M and Christoph Lange (2010). *The Fundamentals of Modern Statistical Genetics*. Springer Science & Business Media.

Lambert, Christophe G and Laura J Black (2012). "Learning from our GWAS mistakes: from experimental design to scientific method." *Biostatistics* 13.2, pp. 195–203.

LaMotte, Lynn Roy (2007). "A direct derivation of the REML likelihood function". *Statistical Papers* 48.2, pp. 321–327.

Lander, E S and N J Schork (1994). "Genetic dissection of complex traits." *Science* 265.5181, pp. 2037–48.

Lander, E S, L M Linton, B Birren, C Nusbaum, M C Zody, J Baldwin, K Devon, K Dewar, M Doyle, et al. (2001). "Initial sequencing and analysis of the human genome." *Nature* 409.6822, pp. 860–921.

LANGE, KENNETH, JOAN WESTLAKE, and M Spence (1976). "Extensions to pedigree analysis III. Variance components by the scoring method". *Annals of human genetics* 39.4, pp. 485–491.

Lao, Oscar, Timothy T Lu, Michael Nothnagel, Olaf Junge, Sandra Freitag-Wolf, Amke Caliebe, Miroslava Balascakova, Jaume Bertranpetit, Laurence A Bindoff, et al. (2008). "Correlation between genetic and geographic structure in Europe." *Curr. Biol.* 18.16, pp. 1241–8.

Lathrop, GM, M Farrall, P O'Connell, B Wainwright, M Leppert, Y Nakamura, N Lench, H Kruyer, M Dean, M Park, et al. (1988). "Refined linkage map of chromosome 7 in the region of the cystic fibrosis gene". *American journal of human genetics* 42.1, p. 38.

Ledoit, Olivier and Michael Wolf (2004). "A well-conditioned estimator for large-dimensional covariance matrices". *Journal of multivariate analysis* 88.2, pp. 365–411.

Lee, S H, J Yang, M E Goddard, P M Visscher, and N R Wray (2012a). "Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood." *Bioinformatics* 28.19, pp. 2540–2.

Lee, S Hong, Teresa R DeCandia, Stephan Ripke, Jian Yang, Patrick F Sullivan, Michael E Goddard, et al. (2012b). "Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs." *Nat. Genet.* 44.3, pp. 247–50.

Lee, S Hong, Denise Harold, Dale R Nyholt, Michael E Goddard, Krina T Zondervan, Julie Williams, et al. (2013). "Estimation and partitioning of polygenic variation captured by common SNPs for Alzheimer's disease, multiple sclerosis and endometriosis." *Hum. Mol. Genet.* 22.4, pp. 832–41.

Lee, Sang Hong, Julius H J van der Werf, Ben J Hayes, Michael E Goddard, and Peter M Visscher (2008). "Predicting unobserved phenotypes for complex traits from whole-genome SNP data." *PLoS Genet.* 4.10, e1000231.

Lee, Sang Hong, Michael E Goddard, Peter M Visscher, and Julius Hj van der Werf (2010). "Using the realized relationship matrix to disentangle confounding factors for the estimation of genetic variance components of complex traits." *Genet. Sel. Evol.* 42, p. 22.

Lee, Seunggeun, Fred A Wright, and Fei Zou (2011). "Control of population stratification by correlation-selected principal components." *Biometrics* 67.3, pp. 967–74.

Lee, Seunggeun, Mary J Emond, Michael J Bamshad, Kathleen C Barnes, Mark J Rieder, Deborah A Nickerson, David C Christiani, Mark M Wurfel, and Xihong Lin (2012c). "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies." *Am. J. Hum. Genet.* 91.2, pp. 224–37.

Leek, Jeffrey T, W Evan Johnson, Hilary S Parker, Andrew E Jaffe, and John D Storey (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." *Bioinformatics* 28.6, pp. 882–3.

Lenhard, Boris, Albin Sandelin, and Piero Carninci (2012). "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." *Nat. Rev. Genet.* 13.4, pp. 233–45.

Levy, Samuel, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, et al. (2007). "The diploid genome sequence of an individual human." *PLoS Biol.* 5.10, e254.

Li, Heng (2011). "A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data." *Bioinformatics* 27.21, pp. 2987–93.

Li, Heng and Richard Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." *Bioinformatics* 25.14, pp. 1754–60.

Li, Jun Z, Devin M Absher, Hua Tang, Audrey M Southwick, Amanda M Casto, Sohini Ramachandran, Howard M Cann, Gregory S Barsh, Marcus Feldman, et al. (2008). "Worldwide human relationships inferred from genome-wide patterns of variation." *Science* 319.5866, pp. 1100–4.

Li, Mingyao, Kai Wang, Struan F A Grant, Hakon Hakonarson, and Chun Li (2009). "ATOM: a powerful gene-based association test by combining optimally weighted markers". *Bioinformatics* 25.4, pp. 497–503.

Lin, Dan-Yu and Patrick F Sullivan (2009). "Meta-analysis of genome-wide association studies with overlapping subjects." *Am. J. Hum. Genet.* 85.6, pp. 862–72.

Lin, Xinyi, Seunggeun Lee, David C Christiani, and Xihong Lin (2013). "Test for interactions between a genetic marker set and environment in generalized linear models." *Biostatistics* 14.4, pp. 667–81.

Lin, Xinyi, Seunggeun Lee, Michael C Wu, Chaolong Wang, Han Chen, Zilin Li, and Xihong Lin (2016). "Test for rare variants by environment interactions in sequencing association studies." *Biometrics* 72.1, pp. 156–64.

Lippert, Christoph, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman (2011). "FaST linear mixed models for genome-wide association studies." *Nat. Methods* 8.10, pp. 833–5.

Lippert, Christoph, Jing Xiang, Danilo Horta, Christian Widmer, Carl Kadie, David Heckerman, and Jennifer Listgarten (2014a). "Greater power and computational efficiency for kernel-based association testing of sets of genetic variants." *Bioinformatics* 30.22, pp. 3206–14.

Lippert, Christoph, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle (2014b). "LIMIX: genetic analysis of multiple traits". *bioRxiv*, p. 003905.

Lippert, Christoph, Francesco Paolo Casale, Barbara Rakitsch, and Oliver Stegle (2014c). "LIMIX: genetic analysis of multiple traits". *BioRxiv*, p. 003905.

Listgarten, Jennifer, Christoph Lippert, Eun Yong Kang, Jing Xiang, Carl M Kadie, and David Heckerman (2013). "A powerful and efficient set test for genetic markers that handles confounders." *Bioinformatics* 29.12, pp. 1526–33.

Listgarten, Jennifer, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman (2012). "Improved linear mixed models for genome-wide association studies." *Nat. Methods* 9.6, pp. 525–6.

Liu, Chuanhai, Donald B Rubin, and Ying Nian Wu (1998). "Parameter expansion to accelerate EM: The PX-EM algorithm". *Biometrika* 85.4, pp. 755–770.

Liu, Dong C and Jorge Nocedal (1989). "On the limited memory BFGS method for large scale optimization". *Mathematical programming* 45.1-3, pp. 503–528.

Liu, Jimmy Z, Allan F McRae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, Nicholas K Hayward, Grant W Montgomery, et al. (2010). "A versatile gene-based test for genome-wide association studies." *Am. J. Hum. Genet.* 87.1, pp. 139–45.

Loh, Po-Ru, Gaurav Bhatia, Alexander Gusev, Hilary K Finucane, Brendan K Bulik-Sullivan, Samuela J Pollack, Teresa R de Candia, Sang Hong Lee, et al. (2015a). "Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis." *Nat. Genet.* 47.12, pp. 1385–92.

Loh, Po-Ru, George Tucker, Brendan K Bulik-Sullivan, Bjarni J Vilhjálmsson, Hilary K Finucane, Rany M Salem, Daniel I Chasman, Paul M Ridker, Benjamin M Neale, et al. (2015b). "Efficient Bayesian mixed-model analysis increases association power in large cohorts." *Nat. Genet.* 47.3, pp. 284–90.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biol.* 15.12, p. 550.

Lynch, Michael, Bruce Walsh, et al. (1998). *Genetics and analysis of quantitative traits*. Vol. 1. Sinauer Sunderland, MA.

MacLeod, IM, PJ Bowman, CJ Vander Jagt, M Haile-Mariam, KE Kemper, AJ Chamberlain, C Schrooten, BJ Hayes, and ME Goddard (2016). "Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits". *BMC genomics* 17.1, p. 144.

Maher, Brendan (2008). "Personal genomes: The case of the missing heritability." *Nature* 456.7218, pp. 18–21.

Marchini, Jonathan and Bryan Howie (2010). "Genotype imputation for genome-wide association studies." *Nat. Rev. Genet.* 11.7, pp. 499–511.

Marchini, Jonathan, Lon R Cardon, Michael S Phillips, and Peter Donnelly (2004). "The effects of human population structure on large genetic association studies." *Nat. Genet.* 36.5, pp. 512–7.

Märtens, Kaspar, Johan Hallin, Jonas Warringer, Gianni Liti, and Leopold Parts (2016). "Predicting quantitative traits from genome and phenome with near perfect accuracy". *Nature communications* 7.

Mather, Kenneth (1938). "Crossing-over". *Biological Reviews* 13.3, pp. 252–292.

McCarthy, Mark I, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John P A Ioannidis, and Joel N Hirschhorn (2008). "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." *Nat. Rev. Genet.* 9.5, pp. 356–69.

McClellan, Jon and Mary-Claire King (2010). "Genetic heterogeneity in human disease." *Cell* 141.2, pp. 210–7.

Mendel, G (1866). *Versuche über Pflanzenhybriden [Experiments on plant hybrids], reprinted in J. Krızenecky.*

Metzker, Michael L (2010). "Sequencing technologies - the next generation." *Nat. Rev. Genet.* 11.1, pp. 31–46.

Meyer, Karin (2007). "WOMBAT: a tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML)." *J Zhejiang Univ Sci B* 8.11, pp. 815–21.

Meyer, Karin et al. (2006). "PX× AI: Algorithmics for better convergence in restricted maximum likelihood estimation". *8th World Congress on Genetics Applied to Livestock Production.*

Modinos, Gemma, Conrad Iyegbe, Diana Prata, Margarita Rivera, Matthew J Kempton, Lucia R Valmaggia, Pak C Sham, Jim van Os, and Philip McGuire (2013). "Molecular genetic gene-environment studies using candidate genes in schizophrenia: a systematic review." *Schizophr. Res.* 150.2-3, pp. 356–65.

Molenberghs, Geert and Geert Verbeke (2007). "Likelihood ratio, score, and Wald tests in a constrained parameter space". *The American Statistician* 61.1, pp. 22–27.

Morgan, Thomas Hunt (1915). *The mechanism of Mendelian heredity.* Holt.

Morris, Andrew P, Cecilia M Lindgren, Eleftheria Zeggini, Nicholas J Timpson, Timothy M Frayling, Andrew T Hattersley, and Mark I McCarthy (2010). "A powerful approach to sub-phenotype analysis in population-based genetic association studies." *Genet. Epidemiol.* 34.4, pp. 335–43.

Moskvina, Valentina and Karl Michael Schmidt (2008). "On multiple-testing correction in genome-wide association studies." *Genet. Epidemiol.* 32.6, pp. 567–73.

Mukhopadhyay, Indranil, Eleanor Feingold, Daniel E Weeks, and Anbupalam Thalamuthu (2010). "Association tests using kernel-based measures of multi-locus genotype similarity between individuals." *Genet. Epidemiol.* 34.3, pp. 213–21.

Ni, Ting, David L Corcoran, Elizabeth A Rach, Shen Song, Eric P Spana, Yuan Gao, Uwe Ohler, and Jun Zhu (2010). "A paired-end sequencing strategy to map the complex landscape of transcription initiation." *Nat. Methods* 7.7, pp. 521–7.

Novembre, John and Matthew Stephens (2008). "Interpreting principal component analyses of spatial population genetic variation." *Nat. Genet.* 40.5, pp. 646–9.

Novembre, John, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, et al. (2008). "Genes mirror geography within Europe." *Nature* 456.7218, pp. 98–101.

Oliehoek, Pieter A, Jack J Windig, Johan A M van Arendonk, and Piter Bijma (2006). "Estimating relatedness between individuals in general populations with a focus on their use in conservation programs." *Genetics* 173.1, pp. 483–96.

Otero, Paola, Emilio Herrera, and Bartolomé Bonet (2002). "Dual effect of glucose on LDL oxidation: dependence on vitamin E." *Free Radic. Biol. Med.* 33.8, pp. 1133–40.

Panoutsopoulou, Kalliope, Konstantinos Hatzikotoulas, Dionysia Kiara Xifara, Vincenza Colonna, Aliki-Eleni Farmaki, Graham RS Ritchie, Lorraine Southam, Arthur Gilly, Ioanna Tachmazidou, Segun Fatumo, et al. (2014). "Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants". *Nature communications* 5.

Pasaniuc, Bogdan and Alkes L Price (2016). "Dissecting the genetics of complex traits using summary association statistics". *bioRxiv*, p. 072934.

Patsopoulos, Nikolaos A, Lisa F Barcellos, Rogier Q Hintzen, Catherine Schaefer, Cornelia M van Duijn, Janelle A Noble, Towfique Raj, et al. (2013). "Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects." *PLoS Genet.* 9.11, e1003926.

Patterson, H Desmond and Robin Thompson (1971). "Recovery of inter-block information when block sizes are unequal". *Biometrika* 58.3, pp. 545–554.

Patterson, Nick, Alkes L Price, and David Reich (2006). "Population structure and eigenanalysis." *PLoS Genet.* 2.12, e190.

Paul, Dirk S and Stephan Beck (2014). "Advances in epigenome-wide association studies for common diseases." *Trends Mol Med* 20.10, pp. 541–3.

Pickrell, Joseph K, Tomaz Berisa, Jimmy Z Liu, Laure Ségurel, Joyce Y Tung, and David A Hinds (2016). "Detection and interpretation of shared genetic influences on 42 human traits." *Nat. Genet.* 48.7, pp. 709–17.

Powell, Joseph E, Peter M Visscher, and Michael E Goddard (2010). "Reconciling the analysis of IBD and IBS in complex trait studies." *Nat. Rev. Genet.* 11.11, pp. 800–5.

Price, Alkes L, Johannah Butler, Nick Patterson, Cristian Capelli, Vincenzo L Pascali, Francesca Scarnicci, Andres Ruiz-Linares, Leif Groop, Angelica A Saetta, et al. (2008). "Discerning the ancestry of European Americans in genetic association studies." *PLoS Genet.* 4.1, e236.

Price, Alkes L, Noah A Zaitlen, David Reich, and Nick Patterson (2010). "New approaches to population stratification in genome-wide association studies." *Nat. Rev. Genet.* 11.7, pp. 459–63.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich (2006). "Principal components analysis corrects for stratification in genome-wide association studies." *Nat. Genet.* 38.8, pp. 904–9.

Price, Alkes L, Agnar Helgason, Gudmar Thorleifsson, Steven A McCarroll, Augustine Kong, and Kari Stefansson (2011). "Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals." *PLoS Genet.* 7.2, e1001317.

Pritchard, J K, M Stephens, N A Rosenberg, and P Donnelly (2000a). "Association mapping in structured populations." *Am. J. Hum. Genet.* 67.1, pp. 170–81.

Pritchard, J K, M Stephens, and P Donnelly (2000b). "Inference of population structure using multilocus genotype data." *Genetics* 155.2, pp. 945–59.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I W de Bakker, et al. (2007). "PLINK: a tool set for whole-genome association and population-based linkage analyses." *Am. J. Hum. Genet.* 81.3, pp. 559–75.

Rakitsch, Barbara, Christoph Lippert, Karsten Borgwardt, and Oliver Stegle (2013). "It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals". *Advances in Neural Information Processing Systems*, pp. 1466–1474.

Rasmussen, Carl Edward (2006). "Gaussian processes for machine learning".

Rietveld, Cornelius A, Sarah E Medland, Jaime Derringer, Jian Yang, Tõnu Esko, Nicolas W Martin, Harm-Jan Westra, Konstantin Shakhbazov, Abdel Abdellaoui, et al. (2013). "GWAS of 126,559 individuals identifies genetic variants associated with educational attainment." *Science* 340.6139, pp. 1467–71.

Ripke, Stephan, Benjamin M Neale, Aiden Corvin, James TR Walters, Kai-How Farh, Peter A Holmans, Phil Lee, Brendan Bulik-Sullivan, David A Collier, Hailiang Huang, et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". *Nature* 511.7510, p. 421.

Risch, N and K Merikangas (1996). "The future of genetic studies of complex human diseases." *Science* 273.5281, pp. 1516–7.

Sabatti, Chiara, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, et al. (2009). "Genome-wide association analysis of metabolic traits in a birth cohort from a founder population." *Nat. Genet.* 41.1, pp. 35–46.

Sasaki, Eriko, Pei Zhang, Susanna Atwell, Dazhe Meng, and Magnus Nordborg (2015). ""Missing" G x E Variation Controls Flowering Time in Arabidopsis thaliana." *PLoS Genet.* 11.10, e1005597.

Saxena, Richa, Marie-France Hivert, Claudia Langenberg, Toshiko Tanaka, James S Pankow, Peter Vollenweider, Valeriya Lyssenko, Nabila Bouatia-Naji, Josée Dupuis, et al. (2010). "Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge." *Nat. Genet.* 42.2, pp. 142–8.

Schadt, Eric E, Stephanie A Monks, Thomas A Drake, Aldons J Lusis, Nam Che, Veronica Colinayo, Thomas G Ruff, Stephen B Milligan, John R Lamb, et al. (2003). "Genetics of gene expression surveyed in maize, mouse and man." *Nature* 422.6929, pp. 297–302.

Schaid, Daniel J, Shannon K McDonnell, Scott J Hebbring, Julie M Cunningham, and Stephen N Thibodeau (2005). "Nonparametric tests of association of multiple genes with human disease." *Am. J. Hum. Genet.* 76.5, pp. 780–93.

Schellenberg, G D, M A Pericak-Vance, E M Wijsman, D K Moore, P C Gaskell, L A Yamaoka, J L Bebout, L Anderson, K A Welsh, C M Clark, et al. (1991). "Linkage analysis of familial Alzheimer disease, using chromosome 21 markers." *Am. J. Hum. Genet.* 48.3, pp. 563–83.

Schifano, Elizabeth D, Michael P Epstein, Lawrence F Bielak, Min A Jhun, Sharon L R Kardia, Patricia A Peyser, and Xihong Lin (2012). "SNP set association analysis for familial data." *Genet. Epidemiol.* 36.8, pp. 797–810.

Schölkopf, Bernhard and Alexander J Smola (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Schor, Ignacio E, Jacob F Degner, Dermot Harnett, Enrico Cannavò, Francesco P Casale, Heejung Shim, David A Garfield, Ewan Birney, Matthew Stephens, Oliver Stegle, et al. (2017). "Promoter shape varies across populations and affects promoter evolution and expression noise". *Nature genetics* 49.4, pp. 550–558.

Scott, Laura J, Karen L Mohlke, Lori L Bonnycastle, Cristen J Willer, Yun Li, William L Duren, Michael R Erdos, Heather M Stringham, Peter S Chines, et al. (2007). "A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants." *Science* 316.5829, pp. 1341–5.

Searle, Shayle R (1982). "Matrix algebra useful for statistics (wiley series in probability and statistics)".

Segura, Vincent, Bjarni J Vilhjálmsson, Alexander Platt, Arthur Korte, Ümit Seren, Quan Long, and Magnus Nordborg (2012). "An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations." *Nat. Genet.* 44.7, pp. 825–30.

Self, Steven G and Kung-Yee Liang (1987). "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions". *Journal of the American Statistical Association* 82.398, pp. 605–610.

Shiraki, Toshiyuki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, et al. (2003). "Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage." *Proc. Natl. Acad. Sci. U.S.A.* 100.26, pp. 15776–81.

Sidore, Carlo, Fabio Busonero, Andrea Maschio, Eleonora Porcu, Silvia Naitza, Magdalena Zoledziewska, Antonella Mulas, Giorgio Pistis, Maristella Steri, Fabrice Danjou, et al. (2015). "Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers". *Nature genetics* 47.11, pp. 1272–1281.

Sikorska, Karolina, Fernando Rivadeneira, Patrick J F Groenen, Albert Hofman, André G Uitterlinden, Paul H C Eilers, and Emmanuel Lesaffre (2013). "Fast linear mixed model computations for genome-wide association studies with longitudinal data." *Stat Med* 32.1, pp. 165–80.

Sivakumaran, Shanya, Felix Agakov, Evropi Theodoratou, James G Prendergast, Lina Zgaga, Teri Manolio, Igor Rudan, Paul McKeigue, James F Wilson, and Harry Campbell (2011). "Abundant pleiotropy in human complex diseases and traits." *Am. J. Hum. Genet.* 89.5, pp. 607–18.

Slatkin, Montgomery (2008). "Linkage disequilibrium–understanding the evolutionary past and mapping the medical future." *Nat. Rev. Genet.* 9.6, pp. 477–85.

Smith, Erin N and Leonid Kruglyak (2008). "Gene-environment interaction in yeast gene expression." *PLoS Biol.* 6.4, e83.

Spector, Tim (2012). *Identically different: why you can change your genes.* Hachette UK.

Speed, Doug, Gibran Hemani, Michael R Johnson, and David J Balding (2012). "Improved heritability estimation from genome-wide SNPs." *Am. J. Hum. Genet.* 91.6, pp. 1011–21.

Speed, Doug, Na Cai, Michael Johnson, Sergey Nejentsev, David Balding, UCLEB Consortium, et al. (2016). "Re-evaluation of SNP heritability in complex human traits". *bioRxiv*, p. 074310.

Stegle, Oliver, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin (2012). "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses." *Nat Protoc* 7.3, pp. 500–7.

Sudlow, Cathie, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, et al. (2015). "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." *PLoS Med.* 12.3, e1001779.

Sudmant, Peter H, Tobias Rausch, Eugene J Gardner, Robert E Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, Goo Jun, et al. (2015). "An integrated map of structural variation in 2,504 human genomes." *Nature* 526.7571, pp. 75–81.

Sul, Jae Hoon, Towfique Raj, Simone de Jong, Paul I W de Bakker, Soumya Raychaudhuri, Roel A Ophoff, Barbara E Stranger, Eleazar Eskin, and Buhm Han (2015). "Accurate and fast multiple-testing correction in eQTL studies." *Am. J. Hum. Genet.* 96.6, pp. 857–68.

Sul, Jae Hoon, Buhm Han, Chun Ye, Ted Choi, and Eleazar Eskin (2013). "Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches." *PLoS Genet.* 9.6, e1003491.

Sun, Wenguang and T Tony Cai (2009). "Large-scale multiple testing under dependence". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2, pp. 393–424.

Teslovich, Tanya M, Kiran Musunuru, Albert V Smith, Andrew C Edmondson, Ioannis M Stylianou, Masahiro Koseki, James P Pirruccello, Samuli Ripatti, Daniel I Chasman, et al. (2010). "Biological, clinical and population relevance of 95 loci for blood lipids." *Nature* 466.7307, pp. 707–13.

The Rat Genome Sequencing and Mapping Consortium (2013). "Combined sequence-based and genetic mapping analysis of complex traits in outbred rats". *Nature genetics* 45.7, pp. 767–775.

Thompson, Robin (1973). "The estimation of variance and covariance components with an application when records are subject to culling". *Biometrics*, pp. 527–550.

Tian, Chao, Robert M Plenge, Michael Ransom, Annette Lee, Pablo Villoslada, Carlo Selmi, Lars Klareskog, Ann E Pulver, Lihong Qi, et al. (2008). "Analysis and application of European genetic substructure using 300 K SNP information." *PLoS Genet.* 4.1, e4.

Trynka, Gosia, Karen A Hunt, Nicholas A Bockett, Jihane Romanos, Vanisha Mistry, Agata Szperl, Sjoerd F Bakker, Maria Teresa Bardella, Leena Bhaw-Rosun, et al. (2011). "Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease." *Nat. Genet.* 43.12, pp. 1193–201.

Tzeng, Jung-Ying, Daowen Zhang, Sheng-Mao Chang, Duncan C Thomas, and Marie Davidian (2009). "Gene-trait similarity regression for multimarker-based association analysis." *Biometrics* 65.3, pp. 822–32.

Tzeng, Jung-Ying, B Devlin, Larry Wasserman, and Kathryn Roeder (2003). "On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit." *Am. J. Hum. Genet.* 72.4, pp. 891–902.

Tzeng, Jung-Ying, Daowen Zhang, Monnat Pongpanich, Chris Smith, Mark I McCarthy, Michèle M Sale, Bradford B Worrall, Fang-Chi Hsu, Duncan C Thomas, and Patrick F Sullivan (2011). "Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression." *Am. J. Hum. Genet.* 89.2, pp. 277–88.

UK10K Consortium et al. (2015). "The UK10K project identifies rare variants in health and disease". *Nature* 526.7571, pp. 82–90.

Valdar, William, Leah C Solberg, Dominique Gauguier, William O Cookson, J Nicholas P Rawlins, Richard Mott, and Jonathan Flint (2006). "Genetic and environmental effects on complex traits in mice." *Genetics* 174.2, pp. 959–84.

VanRaden, P M (2008). "Efficient methods to compute genomic predictions." *J. Dairy Sci.* 91.11, pp. 4414–23.

Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, et al. (2001). "The sequence of the human genome." *Science* 291.5507, pp. 1304–51.

Visscher, Peter M, Sarah E Medland, Manuel A R Ferreira, Katherine I Morley, Gu Zhu, Belinda K Cornes, Grant W Montgomery, and Nicholas G Martin (2006). "Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings." *PLoS Genet.* 2.3, e41.

Visscher, Peter M, Matthew A Brown, Mark I McCarthy, and Jian Yang (2012). "Five years of GWAS discovery." *Am. J. Hum. Genet.* 90.1, pp. 7–24.

Visscher, Peter M, Stuart Macgregor, Beben Benyamin, Gu Zhu, Scott Gordon, Sarah Medland, William G Hill, Jouke-Jan Hottenga, Gonneke Willemsen, et al. (2007). "Genome partitioning of genetic variation for height from 11,214 sibling pairs." *Am. J. Hum. Genet.* 81.5, pp. 1104–10.

Visscher, Peter M, William G Hill, and Naomi R Wray (2008). "Heritability in the genomics era–concepts and misconceptions." *Nat. Rev. Genet.* 9.4, pp. 255–66.

Voight, Benjamin F and Jonathan K Pritchard (2005). "Confounding from cryptic relatedness in case-control association studies." *PLoS Genet.* 1.3, e32.

Wallace, Chris (2013). "Statistical testing of shared genetic control for potentially related traits." *Genet. Epidemiol.* 37.8, pp. 802–13.

Wallace, Chris, Maxime Rotival, Jason D Cooper, Catherine M Rice, Jennie H M Yang, Mhairi McNeill, Deborah J Smyth, David Niblett, François Cambien, et al. (2012). "Statistical colocalization of monocyte gene expression and genetic risk variants for type 1 diabetes." *Hum. Mol. Genet.* 21.12, pp. 2815–24.

Warren, Liling L, Li Li, Matthew R Nelson, Margaret G Ehm, Judong Shen, Dana J Fraser, Jennifer L Aponte, Keith L Nangle, Andrew J Slater, et al. (2012). "Deep resequencing unveils genetic architecture of ADIPOQ and identifies a novel low-frequency variant strongly associated with adiponectin variation." *Diabetes* 61.5, pp. 1297–301.

Waszak, Sebastian M, Olivier Delaneau, Andreas R Gschwind, Helena Kilpinen, Sunil K Raghav, Robert M Witwicki, Andrea Orioli, Michael Wiederkehr, Nikolaos I Panousis, et al. (2015). "Population Variation and Genetic Control of Modular Chromatin Architecture in Humans." *Cell* 162.5, pp. 1039–50.

Wessel, Jennifer and Nicholas J Schork (2006). "Generalized genomic distance-based regression methodology for multilocus association analysis." *Am. J. Hum. Genet.* 79.5, pp. 792–806.

Westfall, Peter H, S Stanley Young, and S Paul Wright (1993). "On adjusting P-values for multiplicity". *Biometrics* 49.3, pp. 941–945.

Wilks, Samuel S (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses". *The Annals of Mathematical Statistics* 9.1, pp. 60–62.

Wilson, Andrew, Elad Gilboa, John P Cunningham, and Arye Nehorai (2014). "Fast kernel learning for multidimensional pattern extrapolation". *Advances in Neural Information Processing Systems*, pp. 3626–3634.

Winkler, Thomas W, Anne E Justice, Mariaelisa Graff, Llilda Barata, Mary F Feitosa, Su Chu, Jacek Czajkowski, Tõnu Esko, Tove Fall, et al. (2015). "The Influence of Age and Sex on Genetic Associations with Adult Body Size and Shape: A Large-Scale Genome-Wide Interaction Study." *PLoS Genet.* 11.10, e1005378.

Wit, H de, W H Dokter, S B Koopmans, C Lummen, M van der Leij, J W Smit, and E Vellenga (1998). "Regulation of p100 (NFKB2) expression in human monocytes in response to inflammatory mediators and lymphokines." *Leukemia* 12.3, pp. 363–70.

Wood, Andrew R, Dena G Hernandez, Michael A Nalls, Hanieh Yaghootkar, J Raphael Gibbs, Lorna W Harries, Sean Chong, Matthew Moore, Michael N Weedon, et al. (2011). "Allelic heterogeneity and more detailed analyses of known loci explain

additional phenotypic variation and reveal complex patterns of association." *Hum. Mol. Genet.* 20.20, pp. 4082–92.

Wood, Andrew R, Tonu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian'an Luan, et al. (2014). "Defining the role of common variation in the genomic and biological architecture of adult human height." *Nat. Genet.* 46.11, pp. 1173–86.

Woodbury, Max A (1950). "Inverting modified matrices". *Memorandum report* 42, p. 106.

Wu, Michael C, Peter Kraft, Michael P Epstein, Deanne M Taylor, Stephen J Chanock, David J Hunter, and Xihong Lin (2010). "Powerful SNP-set analysis for case-control genome-wide association studies." *Am. J. Hum. Genet.* 86.6, pp. 929–42.

Wu, Michael C, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin (2011). "Rare-variant association testing for sequencing data with the sequence kernel association test." *Am. J. Hum. Genet.* 89.1, pp. 82–93.

Xu, ChangJiang, Ioanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini, and Celia M T Greenwood and (2014). "Estimating genome-wide significance for whole-genome sequencing studies." *Genet. Epidemiol.* 38.4, pp. 281–90.

Yang, Jian, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." *Nat. Genet.* 42.7, pp. 565–9.

Yang, Jian, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Pamela A F Madden, Andrew C Heath, Nicholas G Martin, et al. (2012). "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits." *Nat. Genet.* 44.4, 369–75, S1–3.

Yang, Jian, S Hong Lee, Michael E Goddard, and Peter M Visscher (2011a). "GCTA: a tool for genome-wide complex trait analysis." *Am. J. Hum. Genet.* 88.1, pp. 76–82.

Yang, Jian, Teri A Manolio, Louis R Pasquale, Eric Boerwinkle, Neil Caporaso, Julie M Cunningham, Mariza de Andrade, Bjarke Feenstra, Eleanor Feingold, et al.

(2011b). "Genome partitioning of genetic variation for complex traits using common SNPs." *Nat. Genet.* 43.6, pp. 519–25.

Yang, Jian, Michael N Weedon, Shaun Purcell, Guillaume Lettre, Karol Estrada, Cristen J Willer, Albert V Smith, Erik Ingelsson, Jeffrey R O'Connell, et al. (2011c). "Genomic inflation factors under polygenic inheritance." *Eur. J. Hum. Genet.* 19.7, pp. 807–12.

Young, Alexander I, Fabian Wauthier, and Peter Donnelly (2016). "Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index." *Nat Commun* 7, p. 12724.

Yu, Jianming, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, et al. (2006). "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness." *Nat. Genet.* 38.2, pp. 203–8.

Zhao, Guolin, Rachel Marceau, Daowen Zhang, and Jung-Ying Tzeng (2015). "Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression." *Genetics* 199.3, pp. 695–710.

Zhao, Keyan, María José Aranzana, Sung Kim, Clare Lister, Chikako Shindo, Chunlao Tang, Christopher Toomajian, Honggang Zheng, Caroline Dean, et al. (2007). "An Arabidopsis example of association mapping in structured samples." *PLoS Genet.* 3.1, e4.

Zhou, Minghai, Gregory Ottenberg, Gian Franco Sferrazza, and Corinne Ida Lasmézas (2012). "Highly neurotoxic monomeric α-helical prion protein." *Proc. Natl. Acad. Sci. U.S.A.* 109.8, pp. 3113–8.

Zhou, Xiang and Matthew Stephens (2014). "Efficient multivariate linear mixed model algorithms for genome-wide association studies." *Nat. Methods* 11.4, pp. 407–9.

Zhou, Xiang and Matthew Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies." *Nat. Genet.* 44.7, pp. 821–4.

Zhu, Ciyou, Richard H Byrd, Peihuang Lu, and Jorge Nocedal (1997). "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization". *ACM Transactions on Mathematical Software (TOMS)* 23.4, pp. 550–560.

Zvelebil, Marketa and Jeremy Baum (2007). *Understanding bioinformatics*. Garland Science.